Déchiffrer le code histone : épigénétique et toxicologie placentaire

Raphaël Bilgraer

Université Paris Descartes

École doctorale

Médicament, Toxicologie, Chimie, Imageries

THÈSE

Présentée le 16 décembre 2014 en vue de l'obtention du grade de

Docteur de l'Université Paris Descartes

Spécialité : Toxicologie

Par

Raphaël Bilgraer

Déchiffrer le code histone : épigénétique et toxicologie placentaire

Composition du jury :

Pr. Richard COLE	Rapporteur
Dr. Julia CHAMOT-ROOKE	Rapporteur
Pr. Robert BAROUKI	Examinateur
Pr. Douglas RUTLEDGE	Examinateur
Dr. Sylvie GILLET	Examinateur
Pr. Olivier LAPREVOTE	Directeur de thèse

“Without data

you’re just another person

with an opinion”

William Edwards Deming

Remerciements

Ce travail de thèse a été réalisé au sein de l'équipe de Chimie-Toxicologie Analytique et Cellulaire (C-TAC) de l'UMR CNRS 8638 COMETE.

Je tiens tout d'abord à remercier le **Professeur Olivier Laprévote** qui m'a permis de découvrir l'univers de la recherche scientifique lorsque j'étais étudiant en M1 et m'a donné depuis la chance d'évoluer sereinement dans un environnement si enrichissant.

Je remercie tout particulièrement le **Docteur Sylvie Gillet** pour son encadrement au quotidien, sa confiance, son regard critique mais toujours bienveillant, ainsi que pour avoir su me transmettre la rigueur et la remise en question nécessaire à tout scientifique.

J'adresse mes sincères remerciements au **Professeur Richard Cole** ainsi qu'au **Docteur Julia Chamot-Rooke** pour avoir accepté d'être rapporteurs de ce travail de thèse et pour le temps précieux qu'ils ont bien voulu consacrer à cette tâche.

Je suis également profondément reconnaissant aux **Professeurs Robert Barouki** et **Douglas Rutledge,** examinateurs, pour avoir accepté d'évaluer ce travail ainsi que pour m'avoir transmis au cours de mon cursus des connaissances théoriques dans leur domaine nécessaires au bon déroulement de ma thèse.

Un très grand merci à tous les membres permanents du laboratoire que j'ai eu la chance de côtoyer au quotidien, **Martine Radionoff**, **Emmanuel Bourgogne**, **Jean-Michel Warnet**, **France Massicot**, **Nicolas Auzeil**, **Patrice Rat, Anne Regazzetti**, **Delphine Dargère**, **Pascale Leproux**, **Mélodie Dutot**, pour leur bienveillance et leurs attentions à mon égard.

Merci également à **Françoise Baudouin**, **Sophie Achard**, **Marie André**, **Hélène Chapy** et **Marie Gence** pour tous ces agréables moments passés aux travaux pratiques de toxicologie.

Enfin, comment ne pas remercier tous les étudiants du laboratoire, Abdellah, Justine, Julia, Anaïs, Elodie, Nouzha, Kamélia, pour la solidarité, la complicité et la bonne humeur que nous avons su créer tous ensemble.

Sommaire

Liste des principales abréviations

ACN : acétonitrile

ACP : analyse en composantes principales

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

ARNm : ARN messager

ARNmi : micro-ARN

ARNnc : ARN non-codants

B[*a*]P : benzo[*a*]pyrène

CpG : dinucléotide cytosine – guanine

CYP : cytochrome P450

DNMT : ADN méthyltransférase

EIC : chromatogramme reconstitué

ESI : électronébulisation

HAP : hydrocarbure aromatique polycyclique

HAT : histone-acétyltransférase

HDAC : histone-désacétylase

HDM : histone-déméthylase

HMT : histone-méthyltransférase

HPLC : chromatographie liquide haute-performance

MALDI : désorption-ionisation laser assistée par matrice

Q : analyseur quadripôlaire

SAM : S-adénosyl-méthionine

SDS-PAGE : électrophorèse sur gel de polyacrylamide en conditions dénaturantes

SVF : sérum de veau fœtal

TET : famille *ten-eleven translocation* d'enzymes à activité ADN méthyltransférase

TOF : analyseur à temps de vol

UPLC : chromatographie liquide ultra-performance

Liste des Figures

Liste des Tableaux

Partie 1

INTRODUCTION BIBLIOGRAPHIQUE

Chapitre I

Généralités

Chapitre II

Épigénétique et régulation de l'expression des gènes

Chapitre III

Épigénétique, environnement et toxicologie

Chapitre IV

Méthodes analytiques pour déchiffrer le code histone

I. Généralités

I.1 De la génétique à la biologie du développement

Il a fallu attendre la deuxième moitié du XIX^{e} siècle et les travaux de Gregor Mendel pour dissocier la transmission héréditaire des caractères et la biologie du développement. Après avoir été volontairement ignorées par ses pairs pendant près de 35 ans car jugées trop révolutionnaires, ce n'est qu'en 1900 que les lois de Mendel ont été redécouvertes pour donner naissance à la génétique telle que nous la connaissons aujourd'hui.

Thomas Hunt Morgan, célèbre embryologiste américain, faisait partie des pionniers de cette génétique moderne. Avant même la découverte du support de l'information génétique par James Watson et Francis Crick en 1953[1], il s'interrogeait déjà sur l'existence éventuelle de phénomènes capables d'influencer l'expression des gènes. À travers sa célèbre interrogation «si les caractères de l'individu sont déterminés par les gènes, pourquoi toutes les cellules d'un organisme ne sont-elles pas identiques ?», il bouleversait sans le savoir la théorie fondamentale de la biologie moléculaire. Mais ce n'est finalement que vers le milieu du XX^{e} siècle que des éminents biologistes ont entrepris de recréer des liens entre la génétique et la biologie du développement au travers de ce qui allait devenir une seule et même discipline : l'épigénétique.

Venant du mot épigenèse, théorie selon laquelle l'embryon est indifférencié lors des phases précoces du développement, le terme épigénétique a été proposé pour la première fois par Conrad H. Waddington en 1942. Il définissait alors l'ensemble des évènements responsables de la mise en place du programme génétique[2]. Au fil des réflexions scientifiques contemporaines, l'épigénétique a fini par englober tous les phénomènes héréditaires que la génétique était incapable d'expliquer. Cependant, chacun en avait sa propre définition et aucun mécanisme spécifique n'était proposé.

I.2 La naissance des mécanismes épigénétiques

L'importance des travaux de Conrad H. Waddington sur la drosophile découle du lien établi l'activité des gènes et les phases de développement embryonnaire. Il est ainsi apparu que plusieurs phénomènes survenant au cours du développement embryonnaire restaient sans explication, le premier d'entre eux concernant la différenciation cellulaire. Certaines cellules différenciées ont la faculté de maintenir leur phénotype au fil des divisions cellulaires, suggérant qu'il existe des gènes spécialisés responsables de cette différenciation qui sont activés de façon permanente et, à l'inverse, des gènes spécifiques d'autres types cellulaires qui eux sont inactivés. Le deuxième exemple se rapporte aux cellules souches par définition totipotentes qui sont capables de donner naissance à des cellules différenciées ainsi qu'à une nouvelle cellule souche indifférenciée. Dans cette situation, il y a clairement un changement dynamique de l'activité de certains gènes associé à la division cellulaire. En d'autres termes, tous les gènes ne semblent pas actifs en même temps, et leurs profils d'expression paraissent se transmettre de génération en génération *via* une sorte de mémoire cellulaire. Enfin le dernier exemple concerne l'inactivation du chromosome X chez les femelles de mammifères. Très tôt au cours du développement, un des deux exemplaires du chromosome X est inactivé de manière aléatoire dans toutes les cellules de l'organisme. Leur différence d'activité ne peut être expliquée que par des phénomènes intrinsèques aux chromosomes eux-mêmes. Toutes ces constatations ont abouti à la découverte des mécanismes épigénétiques responsables de la régulation transmissible de l'expression des gènes sans altérer la séquence d'ADN.

II. Épigénétique et régulation de l'expression des gènes

La régulation épigénétique de l'expression des gènes est un processus complexe médié par plusieurs mécanismes cellulaires et moléculaires qui agissent de concert. D'après la base de données UniProtKB/SwissProt, environ 10% des protéines codées par le génome jouent un rôle dans la transcription et sa régulation. Ce contrôle épigénétique peut se faire aussi bien en amont qu'en aval de la transcription et est assuré par divers acteurs : la méthylation et

l'hydroxyméthylation de l'ADN, les histones, leurs variants et leurs modifications post-traductionnelles constituant le code histone, ainsi que certains ARN non-codants. La somme de l'information épigénétique constitue ce que l'on appelle l'épigénome d'une cellule ou d'un tissu, défini à un temps *t* et dans des conditions environnementales déterminées. De plus, la réplication de l'ADN s'accompagnant d'une duplication à l'identique de la chromatine[3], une partie de l'épigénome est transmise de génération en génération, au fil des divisions cellulaires, phénomène que l'on pourrait qualifier de mémoire épigénétique. La chromatine est ainsi porteuse d'une information génétique et épigénétique qui peut être résumée sous le terme de chromatome (figure 1).

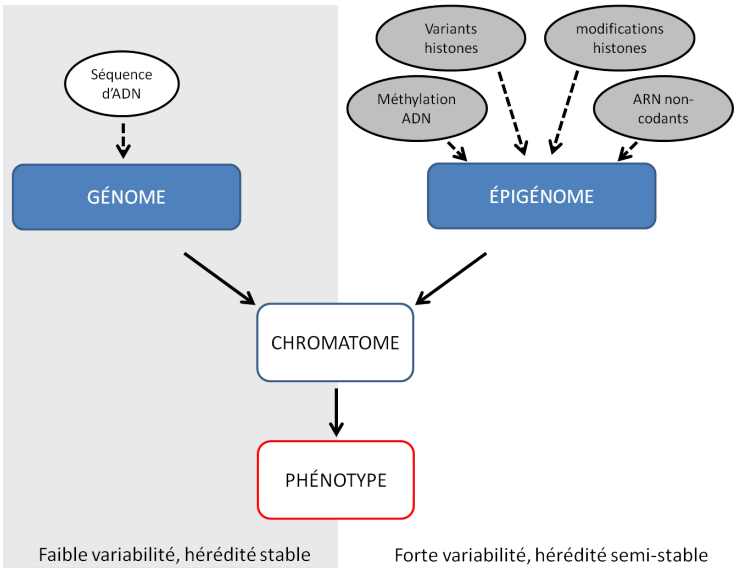


Figure 1 : représentation schématique des facteurs constitutifs du chromatome. Adapté de[4].

Les mécanismes les mieux décrits actuellement sont ceux qui agissent directement sur la chromatine, au niveau pré-transcriptionnel, en modifiant directement ou indirectement sa structure et/ou sa composition. Dans ce chapitre, nous nous intéresserons tout particulièrement aux histones, tout en décrivant brièvement les autres acteurs qui agissent de concert avec elles pour réguler l'expression des gènes.

II.1 La Chromatine

II.1.1 Structure

Chez l'Homme, toutes les cellules d'un même organisme possèdent la même information génétique codée par des molécules d'acide désoxyribonucléique (ADN). La séquence nucléotidique de notre ADN est composée d'environ $3x10^9$ paires de bases. La longueur de cette molécule d'ADN étant plusieurs milliers de fois supérieure au diamètre du noyau des cellules, ces dernières présentent un système de compactage qui associe l'ADN à des protéines structurales, formant ainsi la chromatine[5]. Le nucléosome est défini comme étant l'unité de base de la chromatine et se compose de 147 paires de bases d'ADN qui s'enroulent en une superhélice gauche autour d'un complexe octamérique (figure 2). Le génome d'un mammifère compterait ainsi plus de $1x10^7$ nucléosomes.

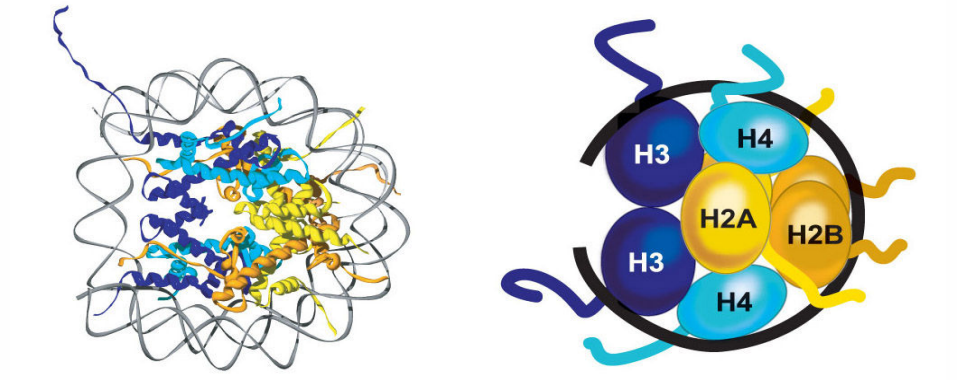


Figure 2 : structure cristallographique d'un nucléosome à 2.8 Å de résolution (gauche) et représentation schématique d'un octamère d'histone autour duquel s'enroule l'ADN double brin (droite). D'après [6].

Ce complexe octamérique se compose de huit sous-unités d'histones de cœur : deux copies de H2A, H2B, H3 et H4. Les nucléosomes sont assemblés et stabilisés *via* l'histone de liaison H1 pour former la structure en collier de perles de la chromatine qui s'enroule et se condense jusqu'à former les boucles de chromatine, puis les chromosomes en début de mitose (figure 3).

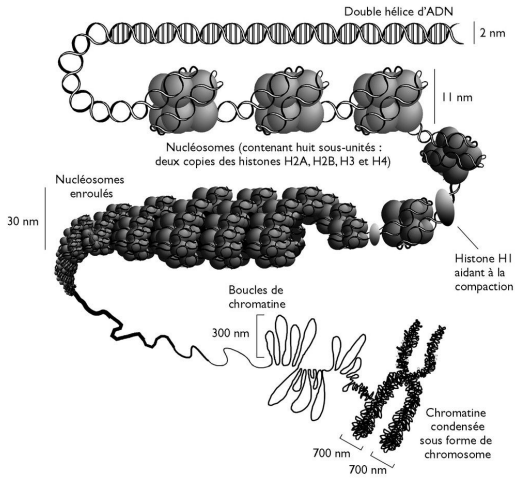


Figure 3 : représentation schématique des différents niveaux de compaction de la chromatine [7]

II.1.2 Variations de l'état de condensation

L'état de condensation de la chromatine n'est pas figé au cours de la vie d'une cellule. Il existe deux types de chromatine qui se distinguent sur la base de critères structuraux et fonctionnels : l'euchromatine et l'hétérochromatine[8]. L'euchromatine correspond à la forme relâchée de la chromatine, c'est-à-dire permissive à la transcription. Elle est particulièrement présente dans des zones riches en gènes dits « actifs », soit environ 4% seulement du génome chez les mammifères. A l'inverse, l'hétérochromatine correspond à une forme condensée de la chromatine qui est répressive à la transcription et majoritairement présente dans des régions non codantes du génome (figure 4). L'état de condensation de la chromatine évolue de façon dynamique par l'intermédiaire de modifications tant de l'ADN que de ses protéines de liaison, notamment les histones. C'est en jouant sur ces modifications de la chromatine que la variabilité cellulaire et tissulaire de l'expression des gènes est assurée.

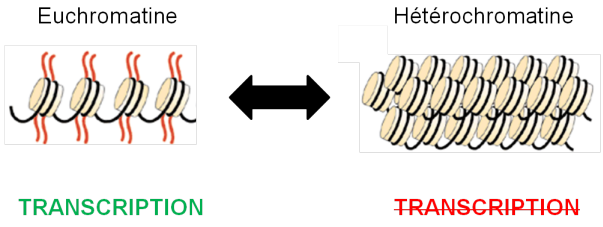


Figure 4 : euchromatine et hétérochromatine. La variation de l'état de condensation de la chromatine influence la transcription.

II.2 La méthylation et l'hydroxyméthylation de l'ADN

La méthylation de l'ADN est le premier mécanisme épigénétique connu permettant à une cellule de rendre certains gènes silencieux. C'est un phénomène réversible qui est impliqué dans de nombreux processus cellulaires (empreinte parentale, inactivation du chromosome X) dès les phases précoces du développement embryonnaire. Largement étudiée depuis les années 1980, il est aujourd'hui clairement établi que la méthylation de l'ADN chez les eucaryotes cible les cytosines, bases pyrimidiques constitutives de l'ADN[9], qui sont alors converties en 5-méthylcytosines (5-mC). Dans la plupart des cas, les cytosines méthylées sont adjacentes à une guanine et forment ensemble un couple de dinucléotides CpG. Les régions du génome enrichies en CpG sont appelées îlots CpG et sont retrouvées préférentiellement au niveau des régions promotrices ou du premier exon de la majorité des gènes[10]. L'ADN méthylé n'est donc pas distribué de manière homogène au sein du génome. Il se concentre au niveau de certaines régions non codantes ou répétitives telles que l'hétérochromatine centromérique ou les transposons[11]. La réaction de méthylation est catalysée par des enzymes appartenant à la famille des méthyltransférases de l'ADN (DNMTs) qui assurent le transfert d'un groupement méthyle (-CH_3) à partir de la S-adénosyl-méthionine (SAM) sur le carbone en position 5 des cytosines (figure 5)[12].

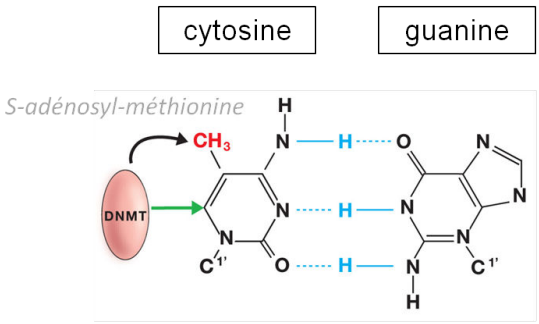


Figure 5 : addition d'un groupement méthyl -CH_3 en position 5 d'une cytosine. Les méthyltransférases de l'ADN s'associent de manière covalente avec le carbone en position 6 pendant le transfert du groupement méthyl depuis une molécule de S-adénosyl-méthionine[6].

La famille des DNMTs comporte plusieurs enzymes qui peuvent soit agir *de novo* en mettant en place le profil initial de méthylation sur la séquence d'ADN, soit

copier après la réplication les profils de méthylation du brin d'ADN parental sur son brin complémentaire.

La méthylation de l'ADN a longtemps été considérée comme la seule marque épigénétique de l'ADN. Ce n'est qu'en 2009 qu'une deuxième modification chimique de l'ADN a été mise en évidence chez l'Homme après avoir déjà été caractérisée chez la bactérie et chez certains mammifères : l'hydroxyméthylation[13]. Cette modification est en réalité le produit d'oxydation des 5-mC, réaction catalysée par des ADN dioxygénases appartenant à un famille d'enzymes codées par les gènes TET (*ten-eleven translocation*)[14]. Les 5-hydroxyméthylcytosines (5-hmC) résultant de cette réaction d'oxydation semblent particulièrement abondantes au niveau du système nerveux central [13] et leur rôle précis n'est pas encore bien compris. Cependant, une richesse en 5-hmC au niveau des promoteurs est souvent associée à une forte activité transcriptionnelle des gènes[15], ces cytosines hydroxyméthylées ayant la capacité de « prendre le dessus » sur les cytosines méthylées. L'hydroxyméthylation des cytosines est également considérée comme la première étape de déméthylation active puisque les 5-hmC peuvent être davantage oxydées par la famille d'enzymes codées par les gènes TET en 5-formylcytosines (5-fC) et en 5-carboxylcytosines (5-caC) qui sont rapidement excisées du génome[16] (figure 6).

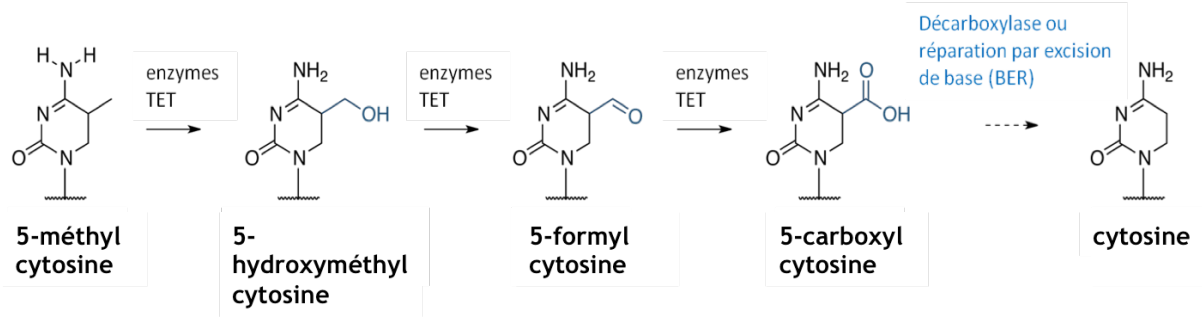


Figure 6 : schéma de la déméthylation active d'une 5-méthylcytosine par oxydations successives suivies d'une décarboxylation ou d'une réparation par excision de base. Adapté de [17].

II.3 Les ARN non-codants

Les ARN non-codants (ARNnc) sont des molécules d'acide ribonucléique (ARN) qui sont issues de la transcription de l'ADN mais qui ne sont pas traduites en protéines. La plupart de ces ARNnc ont un rôle biologique complexe et interagissent avec d'autres molécules d'ARN telles que les ARN messagers (ARNm),

les ARN de transfert (ARNt) ou les ARN ribosomiques (ARNr). Il existe une classe spécifique d'ARNnc particulièrement impliquée dans la régulation de la traduction des ARNm et donc de l'expression des gènes ; celle des ARN interférents (ARNi). Ces ARNi sont de petites molécules d'ARN simple ou double brin qui ont la capacité d'interagir avec un ARNm spécifique et de conduire à sa dégradation ou à la diminution de sa traduction en protéine[18]. Les ARNi sont donc des régulateurs épigénétiques post-transcriptionnels par opposition aux régulateurs pré-transcriptionnels qui agissent sur la chromatine. Il existe trois types majeurs d'ARNi qui diffèrent par leur origine et leur mode d'action : les micro-ARN (ARNmi), les petits ARN interférents (ARNsi) et les ARN interagissant avec Piwi (ARNpi)[19]. Les ARNmi restent à ce jour les ARNnc les plus étudiés d'un point de vue épigénétique, et leur mécanisme d'action semble extrêmement complexe de par la diversité de leurs interactions. Les ARNmi sont des molécules d'ARN composées d'une vingtaine de nucléotides et synthétisées dans le noyau. Initialement, les ARNmi sont composés d'un seul brin qui se replie par la suite pour former une structure double brin non appariée en forme d'épingle à cheveux. Ils sont ensuite clivés dans le cytoplasme par une endoribonucléase de type III, la Dicer, pour former deux brins d'ARNmi distincts. L'un d'eux, généralement le moins stable, se lie au complexe protéique de guidage RISC (*RNA-induced silencing complex*) qui permettra par la suite son appariement avec un ARNm cible de séquence complémentaire, tandis que l'autre sera dégradé. Un ARNmi pourra ainsi diminuer directement l'expression de gènes cibles de deux manières différentes selon la qualité de l'appariement avec son ARNm cible : soit en bloquant la traduction de l'ARNm cible (appariement imparfait), soit en induisant son clivage ou sa dégradation (appariement efficace). Il a également été prouvé dans de récentes études qu'un seul ARNmi pouvait avoir des centaines de cibles différentes[20]. En plus de leur action directe sur la traduction des ARNm, les ARNmi peuvent agir de manière indirecte sur l'expression des gènes. Ils peuvent en effet affecter certains processus de méthylation de l'ADN, notamment lors de la différenciation des cellules souches[21], ou encore réguler d'autres ARNmi.

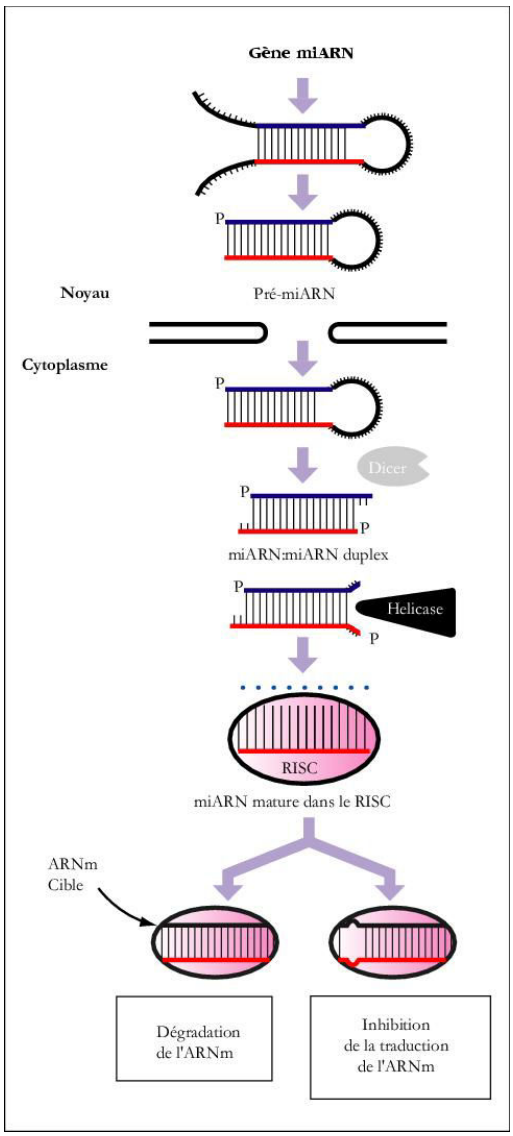


Figure 7 : schéma de la synthèse et du mécanisme d'action directe d'un ARNmi (d'après [22]).

II.4 Les histones et leurs variants

II.4.1 Propriétés biochimiques des histones

Les histones sont de petites protéines nucléaires dont la masse moléculaire des formes non modifiées s'échelonne entre 11 et 17 kilodaltons (kDa) pour les histones de cœur, et atteint 22 kDa pour l'histone de liaison H1. Tous les sous-types d'histones partagent la particularité d'être riches en résidus arginine (R) et lysine (K) ce qui leur confère un caractère hautement basique avec un point isoélectrique (pI) moyen aux alentours de 11. Leur forte interaction avec la molécule d'ADN au sein du nucléosome est assurée par de nombreuses interactions non covalentes, en

particulier les liaisons hydrogène établies entre les acides aminés des histones et les atomes d'oxygène des groupements phosphates de l'ADN, et les interactions électrostatiques entre leurs résidus chargés positivement et les groupements phosphates de l'ADN chargés négativement. Il existe un domaine globulaire hydrophobe de repliement caractéristique des histones appelé *histone-fold*. Ce domaine est composé de trois hélices α reliées entre elles par deux boucles L1 et L2, ainsi que de deux queues flexibles aux extrémités N-terminales (N-ter) et C-terminales (C-ter) dépourvues de structure secondaire (figure 8)[23].

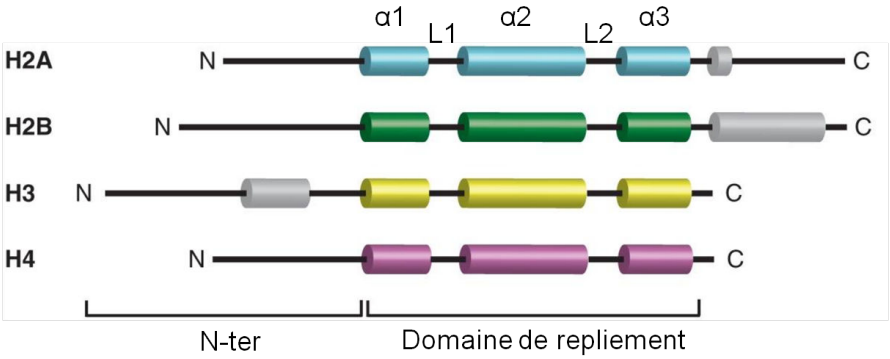


Figure 8 : toutes les histones de cœur partagent le même domaine de repliement (représentée en couleur) qui est composé de trois hélices α. Les hélices α qui sont en dehors de ce domaine sont représentées en gris. Adapté de[24].

Ce domaine commun à toutes les histones assure leur dimérisation selon une configuration dite en poignée de main (figure 9)[25]. Les queues flexibles aux extrémités N-ter et C-ter sont pour leur part impliquées dans les interactions internucléosomiques et peuvent porter certaines modifications post-traductionnelles qui seront détaillées par la suite.

Figure 9 : représentation schématique d'une interaction entre deux domaines histone-fold qui forment un hétérodimère d'histone H3 (jaune) et H4 (violet). D'après [24].

L'histone de liaison H1 présente quant à elle une structure radicalement différente. Elle possède un domaine globulaire hydrophobe GH1 responsable de la liaison à l'ADN. Ce domaine très conservé est composé de trois hélices α et d'un feuillet β en épingle à cheveux en C-ter (figure 10). Les extrémités N-ter et C-ter sont peu conservées et sont impliquées dans de nombreuses interactions protéine-protéine.

Figure 10 : structure du domaine globulaire GH1 de l'histone H1. Les hélices α sont représentées en rouge et le motif en épingle à cheveux en bleu. D'après[26].

II.4.2 Histones canoniques et variants

Les histones font partie des protéines les mieux conservées au cours de l'évolution. Etant codées par de multiples allèles, les histones possèdent de nombreux variants, définis comme étant des isoformes non-alléliques d'une histone canonique qui diffèrent par leurs séquences[27]. Il existe plusieurs variants différents pour chaque sous-type d'histone, excepté pour l'histone H4. Actuellement, ce ne sont pas moins de 55 variants uniques d'histones qui ont été décrits chez l'Homme. Les histones sont ainsi classées en deux groupes majeurs : les histones canoniques et les variants d'histones, encore appelés histones de remplacement. Au-delà de leur séquence en acide aminés, la distinction entre ces deux groupes se fait selon que leur incorporation au sein de la chromatine est dépendante ou non de la réplication de l'ADN[28]. Les histones canoniques sont quasiment exclusivement exprimées au cours de la phase S du cycle cellulaire et incorporées dans la chromatine pour répondre à un besoin important en histones lors de la réplication de l'ADN. Les variants d'histones peuvent eux être synthétisés tout au long du cycle cellulaire et sont la plupart du temps incorporés dans la chromatine

indépendamment de la synthèse d'ADN[29]. Par ailleurs, les gènes qui codent pour les histones canoniques sont dépourvus d'introns, et leurs ARNm ne sont pas polyadénylés mais possèdent sur leur extrémité 3' une structure en tige et boucle[30]. À l'inverse, les pré-ARNm des variants d'histones sont polyadénylés et contiennent des introns et des exons qui peuvent donner lieu à un épissage alternatif[31]. De plus, il est important de noter qu'il existe des isoformes canoniques d'histones H2A, H2B, H3 et H1 qui ne diffèrent parfois que d'un seul acide aminé. Cependant, aucune spécialisation fonctionnelle de ces isoformes canoniques n'a été démontrée à ce jour[32].

L'incorporation des différents variants d'histones de cœur a une répercussion directe sur les interactions entre histones au sein des nucléosomes[33]. Ce phénomène peut influencer la stabilité des nucléosomes et avoir des conséquences sur la conformation de la chromatine, et *a fortiori* sur sa permissivité à la transcription. Certains variants jouent même directement un rôle dans divers processus cellulaires (mitose, réparation de l'ADN, etc.). Par conséquent, le remplacement des histones canoniques par des variants d'histones représente à lui seul un niveau d'information épigénétique faisant partie intégrante du code histone. Nous allons donc détailler l'ensemble des variants caractérisés à ce jour pour chaque sous-type d'histone ainsi que leurs rôles biologiques respectifs.

II.4.2.1 La famille H2A

Parmi les quatre histones de cœur, l'histone H2A est celle qui compte le plus de variants et la plus grande variabilité de séquence, notamment au niveau de son extrémité C-ter. Ainsi, 19 formes différentes de l'histone H2A ont été décrites à ce jour, codées par 26 gènes différents (Tableau 1). L'histone H2A forme un hétérodimère avec l'histone H2B au sein des nucléosomes, et sa forme canonique est composée d'environ 130 acides aminés pour une masse moléculaire moyenne aux alentours de 14 kDa. On ne dénombre pas moins de 11 isoformes canoniques de H2A qui présentent une forte identité de séquence et dont la spécialisation fonctionnelle reste encore à démontrer. En parallèle, il existe 4 variants majeurs de l'histone H2A : H2A.X, H2A.Z, H2A.Bbd et macroH2A. Deux variants mineurs de H2A ont également été caractérisés : H2A.V et H2A.J. Tous ces variants ont des

masses moléculaires significativement différentes des formes canoniques, et certains possèdent également des isoformes.

Le variant H2A.X est un variant majeur qui peut représenter jusqu'à 15% du pool total d'histone H2A. Au niveau de sa séquence, le variant H2A.X diffère des formes canoniques notamment par la présence de quatre acides aminés supplémentaires (SQEL) au niveau de son extrémité C-ter. Il joue un rôle fondamental dans la stabilité du génome *via* son implication dans la signalisation des dommages à l'ADN et le recrutement des complexes de réparation[34]. La présence de ce variant est généralement associée à une conformation ouverte de la chromatine facilitant l'accès aux sites de cassure[35]. La sérine (S) en position 139 de la séquence SQEL est la cible privilégiée de plusieurs enzymes appartenant à la famille des phosphatidylinositol kinases appelées PI3-kinases, dont les kinases ATM (*Ataxia-Telangiectasia-Mutated*), ATR (*Ataxia-Telangiectasia-Related*) et DNA-PK (*DNA-dependant protein kinase*). Lors de cassures double brin de l'ADN, ces kinases vont phosphoryler l'histone H2A.X pour donner naissance à la forme γH2A.X. Ainsi, la forme γH2A.X est-elle particulièrement abondante lors de divers évènements cellulaires tels que la recombinaison méiotique, la fragmentation de l'ADN lors de l'apoptose, la recombinaison V(D)J ou encore sous l'action d'agents génotoxiques (produits chimiques, radiations ionisantes).

Le variant H2A.Z peut, lui, représenter jusqu'à 10% du pool total d'histone H2A. C'est un des variants d'histones les plus conservés chez les eucaryotes supérieurs. Ce variant présente environ 60% d'identité de séquence avec la forme canonique et diffère de celle-ci par la présence au niveau de son extrémité C-ter d'une région acide étendue qui stabilise le nucléosome[36]. Chez les mammifères, H2A.Z est particulièrement présent au niveau des régions promotrices des gènes, et sa présence est corrélée à une activation transcriptionnelle[37]. Par ailleurs, ce variant est particulièrement peu abondant dans les régions hyperméthylées du génome[38]. Enfin, la décompaction de la chromatine induite par l'incorporation de ce variant facilite le recrutement des complexes de réparation de l'ADN au niveau des cassures double brin[39].

Le variant macroH2A est le plus gros variant d'histone avec une masse moléculaire d'environ 40 kDa. Il contient un macrodomaine globulaire en C-ter qui représente les deux-tiers de sa masse moléculaire, et possède deux isoformes,

macroH2A.1 et macroH2A.2, issues d'un épissage alternatif. La présence du variant macroH2A stabilise le nucléosome et induit une compaction relative de la chromatine[40]. Le premier rôle biologique attribué au variant macroH2A est la répression transcriptionnelle, sachant qu'il est particulièrement présent au niveau du chromosome X inactif des femelles de mammifères et qu'il est associé à l'allèle muet lors du phénomène d'empreinte parentale[41]. Plus récemment, certaines études suggèrent que le variant macroH2A pourrait jouer le rôle de capteur en créant un lien entre l'état métabolique d'une cellule et la chromatine[33]. Enfin, macroH2A est surexprimée lors de la différenciation des cellules souches, prévenant ainsi la reprogrammation des cellules souches pluripotentes induites[42].

Contrairement aux autres variants de l'histone H2A, le variant H2A.Bbd (*Barr body-deficient*) est spécifique des mammifères. Ce variant est, contrairement au variant macroH2A, totalement absent du chromosome X inactif des femelles de mammifères et est très abondante au niveau des régions actives du génome[43]. Les nucléosomes contenant H2A.Bbd se révèlent être moins stables que ceux contenant une forme canonique de l'histone H2A et ne contiennent que 118 paires de bases d'ADN contre 147 paires de bases habituellement. Ils confèrent donc à la chromatine une conformation plus relâchée (euchromatine) la rendant permissive à la transcription[44]. D'un point de vue biochimique, ce variant a la particularité de ne contenir qu'une seule lysine dans sa séquence d'acides aminés, le privant ainsi de certaines modifications post-traductionnelles, notamment les acétylations, qui seront vues plus en détail dans un prochain chapitre.

Tableau 1 : récapitulatif des différents variants et isoformes canoniques de la famille d'histone H2A caractérisés chez l'Homme. D'après HIstome[45].

Variants d'histone H2A	N° d'accès UniprotKB	Gènes codants	Masse moyenne (Da)	Identité de séquence avec la forme canonique H2A type 1 (%)
H2A type 1	P0C0S8	HIST1H2AI, HIST1H2AK, HIST1H2AL, HIST1H2AM, HIST1H2AG	13960,29	100
H2A type 1-A	Q96QV6	HIST1H2AA	14102,32	90,8
H2A type 1-B/E	P04908	HIST1H2AE, HIST1H2AB	14004,3	98,5
H2A type 1-C	Q93077	HIST1H2AC	13974,28	98,5
H2A type 1-D	P20671	HIST1H2AD	13976,29	99,2
H2A type 1-H	Q96KK5	HIST1H2AH	13775,06	98,5
H2A type 1-J	Q99878	HIST1H2AJ	13805,09	97,7
H2A type 2-A	Q6FI13	HIST2H2AA4, HIST2H2AA3	13964,3	98,5
H2A type 2-B	Q8IUE6	HIST2H2AB	13864,11	94,6

H2A type 2-C	Q16777	HIST2H2AC	13857,18	93,8
H2A type 3	Q7L7L0	HIST3H2A	13990,28	97,7
macroH2A.1	O75367	H2AFY	39617,06	20,5
macroH2A.2	Q9P0M6	H2AFY2	40058,23	20,8
H2A.Bbd type 1	P0C5Y9	H2AFB1	12697,34	35,1
H2A.Bbd type 2/3	P0C5Z0	H2AFB2, H2AFB3	12713,38	35,8
H2A.J	Q9BTM1	H2AFJ	13888,22	93,1
H2A.V	Q71UI9	H2AFV	13508,69	57,1
H2A.X	P16104	H2AFX	15013,38	82,6
H2A.Z	P0C0S5	H2AFZ	13421,55	56,4

II.4.2.2 La famille H2B

La famille d'histones H2B possède 17 variants caractérisés chez l'Homme, codés par 23 gènes différents (Tableau 2). Ce sont en réalité des isoformes et non des histones de remplacement à proprement parler. Il n'y a donc pas de forme canonique de référence, et le rôle biologique de chacune des isoformes n'a été que très peu étudié jusqu'à aujourd'hui, probablement à cause de leur très forte identité de séquence qui les rend difficilement distinguables. Il n'existe qu'un variant tissu-spécifique : le variant H2B type 1-A ou hTSH2B, qui est uniquement retrouvé au niveau des gonades mâles, et dont le rôle précis reste également à déterminer [46]. Il faut également noter que les isoformes H2B sont nettement moins sujettes aux modifications post-traductionnelles que les autres sous-types d'histone de cœur.

Tableau 2 : récapitulatif des différents variants et isoformes canoniques de la famille d'histone H2B caractérisés chez l'Homme. D'après HIstome[45].

Variants d'histone H2B	N° d'accès UniprotKB	Gènes codants	Masse moyenne (Da)	Identité de séquence avec H2B type 1-B (%)
H2B type 1-B	P33778	HIST1H2BB	13819,00	100
H2B type 1-C/E/F/G/I	P62807	HIST1H2BG, HIST1H2BF, HIST1H2BE, HIST1H2BI, HIST1H2BC	13774,95	97,6
H2B type 1-D	P58876	HIST1H2BD	13804,97	97,6
H2B type 1-H	Q93079	HIST1H2BH	13760,92	96,8
H2B type 1-J	P06899	HIST1H2BJ	13772,97	97,6
H2B type 1-K	O60814	HIST1H2BK	13758,95	96,8
H2B type 1-L	Q99880	HIST1H2BL	13821,02	96,0

H2B type 1-M	Q99879	HIST1H2BM	13858,08	96,8
H2B type 1-N	Q99877	HIST1H2BN	13790,95	98,4
H2B type 1-O	P23527	HIST1H2BO	13774,95	97,6
H2B type 1-A (hTSH2B)	Q96A08	HIST1H2BA	14036,31	85,8
H2B type 2-E	Q16778	HIST2H2BE	13788,97	98,4
H2B type 2-F	Q5QNW6	HIST2H2BF	13788,97	96,0
H2B type 3-B	Q8N257	HIST3H2BB	13776,92	96,0
H2B type F-M	P0C1H6	H2BFM	17001,31	33,8
H2B type F-S	P57053	H2BFS	13813,00	95,2
H2B type W-T	Q7Z2G1	H2BFWT	19618,46	31,4

II.4.2.3 La famille H3

La famille H3 comporte 6 variants différents (Tableau 3). Parmi eux, deux sont des isoformes de l'histone canonique (H3.1 et H3.2) et quatre sont des variants de remplacement (H3.3, CENP-A, H3.3C et H3.1t). Les deux isoformes H3.1 et H3.2 de l'histone canonique ne diffèrent que par un seul acide aminé (cystéine *versus* sérine en position 96). Pourtant, H3.1 est codée par dix gènes différents, alors que H3.2 ne l'est que par un seul. D'un point de vue biologique, ces deux isoformes possèdent des fonctions bien distinctes. L'isoforme H3.1 peut être associée tant à une activation qu'à une répression de la transcription, tandis que H3.2 ne semble être associée qu'à une répression de la transcription des gènes[47]. Concernant les histones H3 de remplacement, il en existe quatre types différents, dont deux sont spécifiques des testicules (H3.1t et H3.3C).

Le variant H3.3 peut représenter jusqu'à 50% du pool total d'histone H3 et est incorporé dans la chromatine lors d'une activation transcriptionnelle[48]. En effet, la présence de H3.3 au sein des nucléosomes confère à la chromatine une conformation relâchée davantage permissive à la transcription qu'avec la forme canonique H3.1[49]. Le variant H3.3 est également co-localisé avec le variant H2A.Z au niveau des promoteurs des gènes actifs[50].

Le variant CENP-A (*centromer protein A*) est tout à fait particulier. Il est appelé variant centromérique de H3 car il est spécifiquement localisé au niveau des régions centromériques des chromosomes où il participe à la formation des kinétochores actifs. Les centromères sont des régions spécifiques de la chromatine

où s'assemblent les kinétochores. Ils servent ainsi de point d'ancrage aux microtubules du fuseau mitotique lors de la division cellulaire[51]. La seule séquence d'ADN centromérique ne suffisant pas pour former des centromères fonctionnels, c'est *via* le mécanisme épigénétique d'incorporation du variant CENP-A que l'identité des centromères est établie[52].

Tableau 3 : récapitulatif des différents variants et isoformes canoniques de la famille d'histone H3 caractérisés chez l'Homme. D'après HIstome[45].

Variants d'histone H3	N° d'accès UniprotKB	Gènes codants	Masse moyenne (Da)	Identité de séquence avec la forme canonique H3.1 (%)
H3.1	P68431	HIST1H3A, HIST1H3D, HIST1H3C, HIST1H3E, HIST1H3I, HIST1H3G, HIST1H3J, HIST1H3H, HIST1H3B, HIST1H3F	15272,89	100
H3.2	Q71DI3	HIST2H3C, HIST2H3A, HIST2H3D	15256,83	99,3
CENP-A	P49450	CENPA	15990,57	44,1
H3.1t	Q16695	HIST3H3	15377,06	97,1
H3.3	P84243	H3F3A, H3F3B	15196,72	96,3
H3.3C	Q6NXT2	H3F3C	15082,54	93,4

II.4.2.4 La famille H1

L'histone H1 dite de liaison est essentielle à la stabilisation des nucléosomes et à la compaction de la chromatine. Elle stabilise l'interaction entre l'octamère d'histones de cœur et l'ADN au sein du nucléosome. Il n'y a ainsi qu'une seule copie de l'histone H1 par nucléosome. D'un point de vue biologique, elle semble jouer un rôle dynamique dans la régulation de la transcription en participant tant à sa répression qu'à son activation[53]. Onze formes différentes de l'histone H1 ont été caractérisées chez l'Homme, avec des séquences qui divergent significativement (Tableau 4). Les variants H1.2 à H1.5 sont exprimés dans toutes les cellules somatiques et retrouvés dans tous les types de tissus. Le variant H1.1 n'est lui retrouvé que dans certains tissus spécifiques (thymus, testicules, rate, lymphocytes). L'expression des variants H1t, H1T2 et HILS1 est spécifique des gonades mâles, et $H1_{oo}$ est quant à elle exprimée exclusivement dans les ovocytes[54]. D'autre part, l'expression des formes H1.1, H1.2, H1.3, H1.4, H1.5 et H1t se fait exclusivement pendant la phase S du cycle cellulaire. Ce sont donc des isoformes canoniques contrairement aux véritables variants H1.0, H1.X, H1T2,

HILS1 et $H1_{oo}$ qui sont exprimés tout au long du cycle cellulaire et incorporés dans la chromatine indépendamment de la réplication de l'ADN[55]. Le variant HILS1 est exprimé lors des phases tardives de la spermatogénèse où il participe à la condensation de la chromatine[56]. Le variant testiculaire H1T2 est impliqué dans l'élongation des spermatides et dans l'échange des histones avec les protamines, suggérant ainsi qu'il participe à la condensation de la chromatine[57]. Le variant ovocyte-spécifique $H1_{oo}$ joue quant à lui un rôle important dans la maturation des ovocytes et est incorporé dans la chromatine des spermatozoïdes après la fertilisation, induisant ainsi sa condensation[58].

Tableau 4 : récapitulatif des différents variants et isoformes canoniques de la famille d'histone de liaison H1 caractérisés chez l'Homme. D'après HIstome[45].

Variants d'histone H1	N° d'accès UniprotKB	Gènes codants	Masse moyenne (Da)	Identité de séquence avec la forme canonique H1.2 (%)
H1.1	Q02539	HIST1H1A	21710,90	64,6
H1.2	P16403	HIST1H1C	21233,56	100
H1.3	P16402	HIST1H1D	22218,71	79,6
H1.4	P10412	HIST1H1E	21734,08	84,1
H1.5	P16401	HIST1H1B	22448,98	72,9
H1.0	P07305	H1F0	20731,74	36,2
H1oo	Q8IZA3	H1FOO	35813,49	19,4
H1t	P22492	HIST1H1T	21887,81	51,8
H1.X	Q92522	H1FX	22355,92	32,2
H1T2	Q75WM6	H1FNT	28115,95	18,7
HILS1	P60008	HILS1	25631,73	16,3

II.4.2.5 L'histone H4

L'histone H4 est un cas particulier. Elle est la plus conservée des histones à travers les espèces et elle ne possède aucun variant ni isoforme identifié à ce jour. Elle existe donc sous une forme unique codée par 14 gènes différents chez l'Homme (Tableau 5). Son rôle biologique varie non pas en fonction de la forme sous laquelle elle est incorporée dans la chromatine mais en fonction des modifications post-traductionnelles qu'elle porte.

Tableau 5 : variant unique de l'histone H4 caractérisé chez l'Homme. D'après HIstome [45].

Variant H4	N° d'accès UniprotKB	Gènes codants	Masse moyenne (Da)
H4	P62805	HIST4H4, HIST2H4B, HIST1H4I, HIST1H4A, HIST1H4D, HIST1H4F, HIST1H4K, HIST1H4J, HIST1H4C, HIST1H4H, HIST1H4B, HIST1H4E, HIST1H4L, HIST2H4A	11236,15

II.4.3 Incorporation des variants d'histones et complexes de remodelage de la chromatine

Au cours de la synthèse de l'ADN, l'assemblage des nucléosomes se déroule grossièrement selon un processus qui démarre par l'incorporation d'un dimère de dimère (H3-H3 et H4-H4) suivi de l'ajout de deux hétérodimères (H2A-H2B). Ce processus est facilité par l'intervention de molécules chaperons qui interagissent avec les histones canoniques. Ainsi, le tétramère $(H3\text{-}H4)_2$ est-il incorporé à l'ADN grâce au chaperon CAF-1 (*Chromatin Assembly Factor 1*), qui est lui-même déposé sur l'ADN en cours de synthèse par la protéine multimérique PCNA (*Proliferating Cell Nuclear Antigen*). L'incorporation des dimères (H2A-H2B) serait quant à elle assurée par le chaperon NAP-1 (*Nucleosome Assembly Protein 1*). Ce processus général concerne l'incorporation des histones canoniques durant la réplication. S'agissant des variants d'histones, leur incorporation dans la chromatine est assurée par une machinerie d'assemblage tout à fait différente constituée de complexes de remodelage. Ces complexes de remodelage de la chromatine (CRC) peuvent contenir jusqu'à douze protéines différentes, et tous contiennent une hélicase ATP-dépendante appartenant à la famille Snf2. Ils sont capables de moduler les interactions entre l'ADN et les octamères d'histones, de déplacer les nucléosomes le long de la molécule d'ADN et surtout de modifier leur composition en variants d'histones[59]. Ces complexes enzymatiques jouent donc un rôle essentiel, et il a d'ailleurs été prouvé qu'une délétion ou une mutation des gènes codant pour les protéines les composant avait pour conséquence une perte de contrôle du cycle cellulaire pouvant aboutir à une apoptose excessive ou à une

tumorigénèse[60]. Actuellement, quatre familles majeures de CRC ont été mises en évidence chez l'Homme, chacune ayant des fonctions spécifiques.

II.4.3.1 La famille SWI/SNF

Chez les mammifères, le principal complexe de remodelage de la chromatine appartenant à la famille SWI/SNF est appelé BAF (*Brg/Brm-Associated Factors*). Il possède cinq sous-unités orthologues de la levure et de multiples sous-unités spécifiques des mammifères. Il est ainsi très polymorphe et sa masse moléculaire peut atteindre 2 000 kDa. La nature des sous-unités qui composent ce complexe BAF varie en fonction des étapes de développement cellulaire et est corrélée à l'expression sélective de gènes cibles[61]. Les complexes de la famille SWI/SNF peuvent déplacer l'octamère d'histones soit en le faisant glisser le long du même brin d'ADN (déplacement en cis) soit en le transférant sur un autre brin d'ADN (déplacement en trans). Ils jouent un rôle fondamental lors de toutes les étapes de développement embryonnaire et semblent posséder autant de rôles biologiques que de combinaisons d'assemblage de leurs sous-unités[62].

II.4.3.2 La famille INO80

Le complexe INO80 peut posséder chez les mammifères des sous-unités communes avec le complexe BAF. Il a une affinité particulière pour les variants H2A.X et H2A.Z de l'histone H2A, et est capable d'évincer certaines histones de cœur des nucléosomes. Ce complexe semble jouer un rôle direct dans la réparation de l'ADN, notamment lors de la survenue de cassures double-brin en facilitant le recrutement des complexes de réparation, en introduisant le variant H2A.X au niveau de la cassure et en le remplaçant par un autre variant non phosphorylé une fois le processus de réparation achevé[63].

II.4.3.3 La famille ISWI

Les complexes de remodelage de la famille ISWI fonctionnent de la même façon que ceux de la famille SWI/SNF. Ils ont comme principale fonction l'activation de la transcription au niveau des régions euchromatiques. Toutefois, ils sont également présents au niveau de certaines régions hétérochromatiques, ce qui

laisse penser qu'ils joueraient un rôle d'initiation et de maintien de la formation de l'hétérochromatine[64]. Certaines études affirment même qu'un défaut de fonctionnement de ces complexes à la suite d'une mutation aurait les mêmes conséquences sur la structure de la chromatine qu'une absence totale d'histone de liaison H1[65]. Le complexe WSTF joue quant à lui un rôle dans la réparation de l'ADN en phosphorylant la tyrosine 142 de l'histone H2A.X. A l'image du complexe BAF, les complexes de la famille ISWI ont de multiples possibilités d'assemblage de leurs sous-unités et autant de fonctions biologiques qui en découlent.

II.4.3.4 La famille CHD

La famille CHD compte neuf complexes de remodelage qui se différencient par la nature de leurs domaines structuraux et leurs fonctions biologiques. CHD1 et CHD2 possèdent un domaine de liaison à l'ADN en C-terminal et sont majoritairement présents au niveau de sites transcriptionnellement actifs où ils s'associent avec des facteurs d'initiation de la transcription. Durant le développement embryonnaire, le complexe CHD1 est essentiel au maintien de la pluripotence des cellules souches en stabilisant la chromatine à l'état ouvert. De plus, CHD1 est responsable de l'intégration du variant H3.3 au sein de la chromatine des cellules du sperme durant l'embryogénèse[66].

Les complexes CHD3 et CHD4 ne possèdent pas de domaine de liaison à l'ADN mais deux motifs « doigt *PHD* » en N-terminal. Ils s'intègrent dans un complexe protéique appelé NuRD qui possède une activité à la fois de remodelage de la chromatine et de désacétylation des histones, et qui semble ainsi exercer une activité de répression de la transcription. NuRD est également impliqué dans la réparation des cassures double-brin de l'ADN et dans la progression du cycle cellulaire. Des études récentes démontrent que la sous-expression de certaines des sous-unités du complexe NuRD durant le vieillissement normal ou prématuré conduit à un changement de la structure de la chromatine et à l'apparition de dommages spontanés de l'ADN[67]. Ainsi, le complexe NuRD joue-t-il le rôle de gardien de la stabilité du génome.

Enfin les complexes CHD5 à CHD9 possèdent un domaine fonctionnel supplémentaire sur leur partie C-terminale. Leur absence conduit à une

dérégulation majeure de la transcription de nombreux gènes clés du développement embryonnaire[68].

II.5 Les modifications post-traductionnelles des histones

II.5.1 Définition

Une modification post-traductionnelle est la modification covalente d'une protéine introduite par l'intermédiaire d'une enzyme et qui intervient après l'étape cytoplasmique de traduction. Chez les eucaryotes, il existe une grande diversité chimique de modifications post-traductionnelles qui peuvent influencer la localisation, la demi-vie ou encore l'activité de la protéine cible dans l'organisme. Dans le cas des histones, la plupart des modifications post-traductionnelles sont réversibles et sont localisées au niveau de l'extrémité N-ter des protéines.

II.5.2 Les modifications constitutives du code histone

II.5.2.1 L'hypothèse du code histone

Comme nous l'avons évoqué brièvement lors d'un chapitre précédent, la combinaison entre les variants d'histones incorporés dans la chromatine et les modifications post-traductionnelles qu'ils portent constituent une information combinatoire complexe résumée sous le terme de « code histone ».

Chez l'Homme, neuf types de modifications post-traductionnelles d'histones ont été décrits dans la littérature et de nombreux sites de modification ont été identifiés pour chacune d'entre elles. Classiquement, ces modifications sont réparties en deux groupes selon leur taille. Le premier groupe inclut les modifications chimiques de petite taille, à savoir l'acétylation, la méthylation (mono-, di- et triméthylation), la phosphorylation et la citrullination. Le deuxième groupe inclut les modifications de taille plus importante telles que l'ubiquitinylation, la sumoylation, la ribosylation ou la biotinylation, mais également la crotonylation. Les cibles de ces modifications sont les résidus lysine, arginine, sérine, thréonine ou tyrosine (figure 11). D'un point de vue biologique,

les modifications les plus étudiées sont l'acétylation, la méthylation, la phosphorylation et l'ubiquitinylation. Chacune de ces modifications peut être considérée individuellement ou collectivement. De manière générale, les modifications post-traductionnelles des histones auront un impact sur la conformation de la chromatine et sur sa permissivité à la transcription ce qui permet de relier la présence d'une modification précise à l'activité de certains gènes. En parallèle, la modification d'un résidu peut influencer directement la survenue d'une modification d'un autre résidu voisin au sein de la même protéine. Certaines modifications seront donc mutuellement exclusives, tandis que d'autres seront à l'inverse positivement corrélées. Cet aspect combinatoire doit obligatoirement être pris en compte lorsque l'on s'intéresse à l'impact de ces modifications sur la transcription des gènes. En effet, elles agissent de concert et non de manière isolée[69], et la présence simultanée de plusieurs modifications différentes sur la même protéine aura un effet sur la transcription qu'il sera difficile de prédire[70]. *In fine*, la diversité des formes d'histones, leur richesse en sites potentiellement modifiables ainsi que la diversité chimique des modifications post-traductionnelles décrites laisse entrevoir le nombre de combinaisons possibles (plus d'un million) et donc la complexité de l'information portée par le code histone.

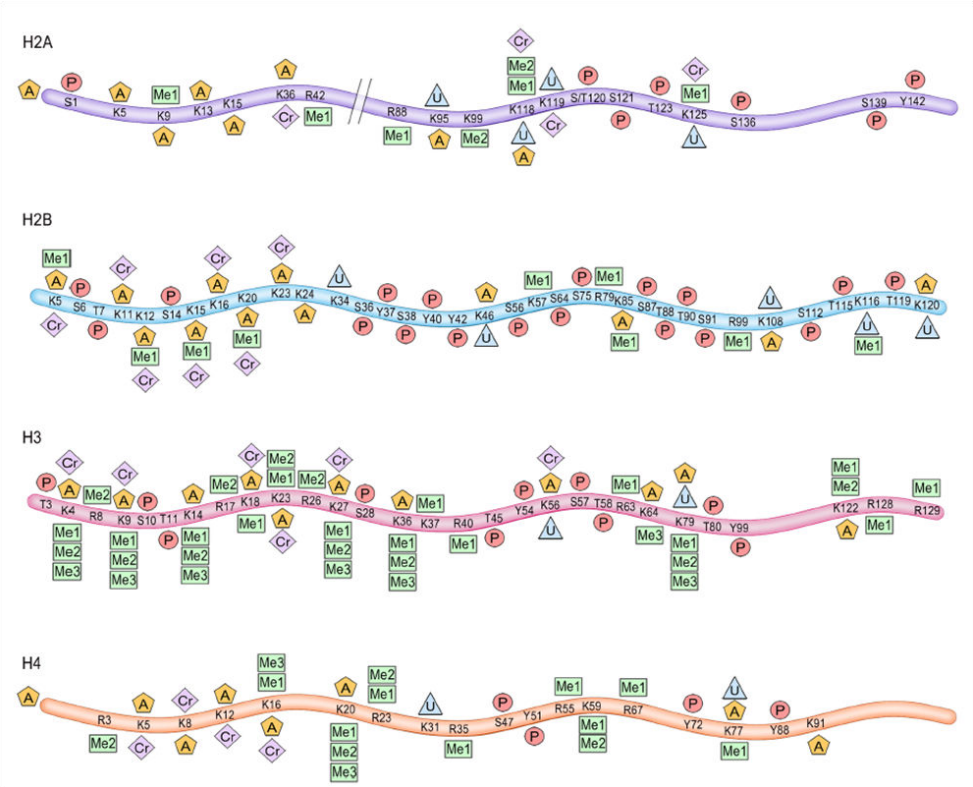


Figure 11 : principales modifications post-traductionnelles affectant les histones de cœur. Chaque type de modification post-traductionnelle est représentée par une couleur et une abréviation différente (A : acétylation, Me1, Me2, Me3 : mono-, di- et triméthylation, P : phosphorylation, U : ubiquitinylation, Cr : crotonylation). D'après [71].

Cette information est déchiffrée *in vivo* par des mécanismes moléculaires qui permettent d'associer à chaque combinaison un état fonctionnel de la chromatine. Le premier niveau de lecture du code histone correspond en effet à la conformation de la chromatine comme nous venons de le décrire. Le second correspond quant à lui à la liaison de transducteurs *via* des interactions protéines-protéines qui peuvent agir comme des inducteurs ou des répresseurs de la transcription et qui feront l'objet d'un prochain chapitre.

II.5.2.2 L'acétylation

L'acétylation des histones est le résultat d'une réaction enzymatique réversible qui correspond au transfert d'un groupement acétyle $-COCH_3$ sur le groupement ε-aminé d'un résidu lysine par attaque nucléophile. Cette réaction de transfert se fait à partir d'une molécule d'acétyl-coenzyme A et est catalysée par des enzymes appartenant à la famille des histone-acétyltransférases (HAT) (figure 12).

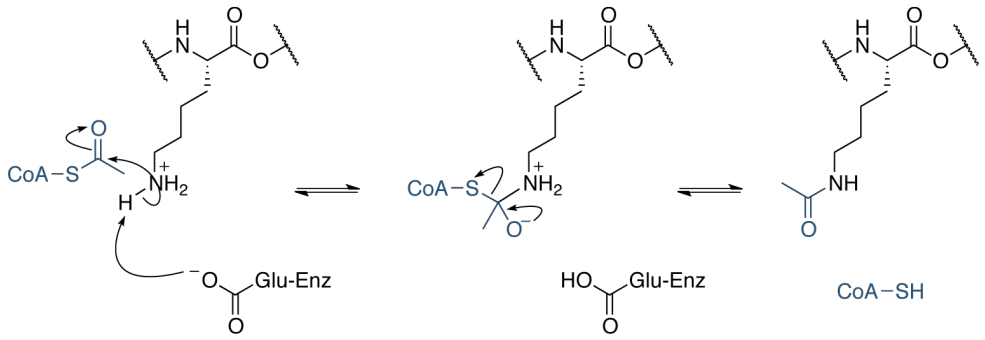


Figure 12 : mécanisme d'acétylation d'une lysine catalysée par les enzymes de la famille des HAT. Ces enzymes possèdent un résidu glutamate qui agit comme une base qui permet de catalyser l'attaque nucléophile du groupement ε-aminé d'un résidu lysine sur le pont thioester de l'acétyl-coenzyme A. La réaction conduit à la libération de coenzyme A (CoA-SH). D'après ATDBio (http://www.atdbio.com/content/56/Epigenetics).

Ces enzymes se trouvent pour la plupart au sein de complexes multiprotéiques qui peuvent assurer d'autres fonctions en parallèle de l'acétylation[72]. Il existe quatre familles de HAT : la superfamille des GNAT (*GCN5-N-Acetyltransferase-related*), la superfamille CBP/p300, la superfamille MYST et la superfamille SRC (*Steroid Receptor Coactivator*). La réaction inverse est catalysée par des enzymes appartenant à la famille des histone-désacétylases (HDAC) dont il existe six classes majeures. Les HDAC appartenant aux classes I, II et IV sont zinc-dépendantes tandis que celles de la classe III, appelées sirtuines, sont NAD-dépendantes.

Il existe donc une balance entre l'activité des HAT et celle des HDAC dont résultera le niveau global d'acétylation des histones. L'acétylation est considérée comme une marque d'activation de la transcription puisque l'addition d'un groupement acétyle neutralise la charge positive de la chaîne latérale des lysines et diminue ainsi les interactions électrostatiques entre les histones et l'ADN chargé négativement. Lors d'une acétylation, la chromatine est donc décondensée et l'ADN est davantage accessible à la machinerie transcriptionnelle[73]. Certaines protéines de remodelage de la chromatine peuvent également se servir des résidus acétylés comme points d'ancrage.

II.5.2.3 La méthylation

La méthylation des histones se fait par transfert d'un groupement méthyle $-CH_3$ à partir d'une molécule de S-adénosine-méthionine (SAM) vers un résidu lysine ou arginine. Cette réaction est catalysée par des enzymes de la famille des histone-méthyltransférases (HMT) dont certaines sont spécifiques des résidus K, les HKMT, ou *Protein lysine methyltransferase*, et les autres des résidus R, les PRMT ou *Protein arginine methyltransferase*. Elles peuvent ajouter entre 1 et 3 groupements méthyle sur le même résidu, aboutissant donc à une mono- ou une diméthylation symétrique ou asymétrique des arginines et à une mono-, une di- ou une triméthylation des lysines (figure 13). Cependant, seules les réactions de mono- et diméthylation des lysines sont réversibles[74], la déméthylation étant catalysée par des enzymes de la famille des histone-déméthylases (HDM) spécifiques des lysines. Contrairement à l'acétylation, les méthylations n'affectent pas la charge globale du résidu modifié mais peuvent par contre ajouter un encombrement stérique et une hydrophobicité réduisant la stabilité du nucléosome. Les résidus méthylés peuvent également servir de point d'ancrage à certaines protéines impliquées dans le remodelage de la chromatine.

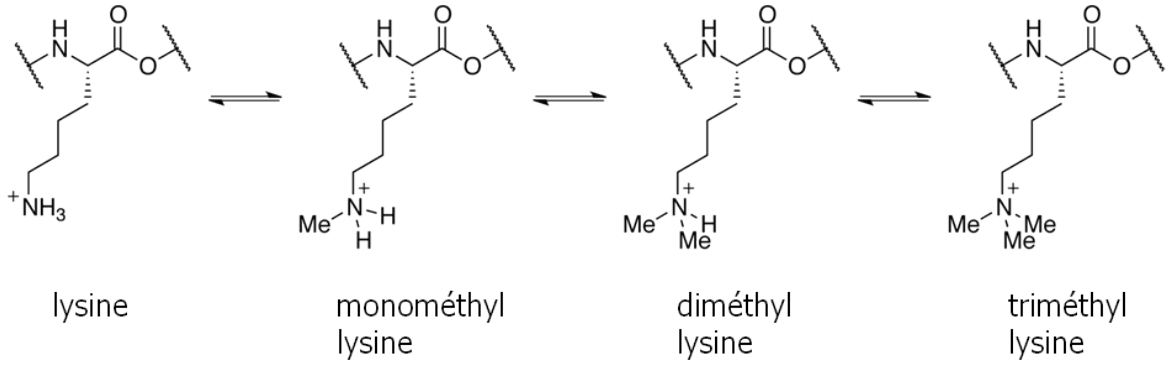


Figure 13 : réaction de mono-, di- et triméthylation des lysines. D'après ATDBio (http://www.atdbio.com/content/56/Epigenetics).

D'un point de vue biologique, le rôle de la méthylation est beaucoup plus complexe que celui de l'acétylation. Tout d'abord, la méthylation d'une même protéine peut avoir deux effets opposés selon le résidu concerné. Par exemple, la monométhylation de la lysine 4 de l'histone H3 (H3K4me) entraînera une activation

de la transcription des gènes tandis que la monométhylation de la lysine 9 (H3K9me) entrainera une répression de la transcription. Ensuite, la méthylation potentielle d'un même résidu à trois degrés différents ajoute un niveau de complexité supplémentaire puisque l'impact sur la chromatine ne sera pas le même.

II.5.2.4 La phosphorylation

Considérant l'ensemble des protéines d'un organisme, la phosphorylation est la plus répandue des modifications post-traductionnelles. Elle correspond au transfert d'un groupement phosphate $-PO_4^{3-}$ à partir d'une molécule d'ATP vers un résidu sérine, thréonine ou tyrosine (figure 14). La réaction de phosphorylation est catalysée par des enzymes de la famille des sérine/thréonine kinases qui ne sont pas spécifiques des histones. L'ajout du groupement phosphate qui possède une charge négative intrinsèque confèrera une charge négative aux résidus cibles, diminuant l'affinité des histones pour l'ADN et induisant ainsi une décondensation localisée de la chromatine. Comme c'est le cas avec les autres protéines, la phosphorylation des histones est une réaction réversible. Cette réaction inverse de déphosphorylation est assurée par des phosphatases non spécifiques des histones.

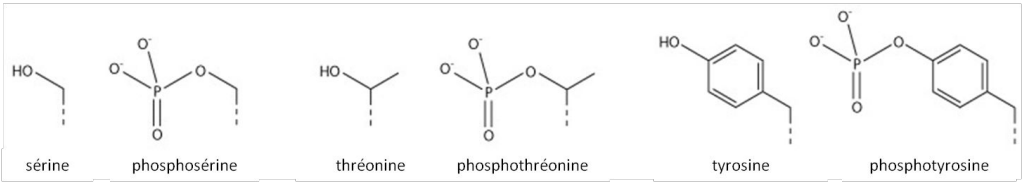


Figure 14 : phosphorylation des résidus sérine, thréonine et tyrosine. (http://commons.wikimedia.org/wiki/File:Amino_acid_phosphorylations.tif#mediaviewer/File:Amino_acid_phosphorylations.tif).

La plupart des phosphorylations sont dépendantes du cycle cellulaire et servent à guider sa progression à travers les différentes phases qui le composent. Un même résidu modifié peut ainsi être impliqué dans différents phénomènes cellulaires en fonction du temps. Les phosphorylations sont également très impliquées dans le « dialogue » entre les différentes modifications post-traductionnelles, notamment

au niveau de l'histone H3. La phosphorylation de la sérine 139 du variant H2A.X est quant à elle classiquement associée aux cassures double-brin de l'ADN et est indispensable au recrutement de la machinerie de réparation[75].

II.5.2.5 Les autres modifications post-traductionnelles

Les autres modifications post-traductionnelles que peuvent subir les histones sont beaucoup moins documentées dans la littérature, mais font l'objet de plus en plus d'études. Il s'agit de l'ubiquitinylation, de la sumoylation, de la ribosylation, de la biotinylation, de la crotonylation, de la citrullination ou encore de la succinylation, de la malonylation et de la carbonylation. Bien que beaucoup moins rencontrées que les modifications précédentes, elles n'en jouent pas moins un rôle biologique significatif. Hormis la citrullination, toutes ces modifications surviennent sur des résidus lysine. L'ubiquitinylation, la biotinylation, la ribosylation et la crotonylation sont réversibles tandis que la sumoylation ne l'est pas. La citrullination est un cas particulier puisqu'il s'agit en réalité d'une réaction de déimination des arginines catalysée par des déiminases PAD (*Peptidyl-arginine deiminase*) (figure 15). Cette réaction de citrullination empêche la méthylation des résidus arginines concernés et peut donc affecter la transcription des gènes[76].

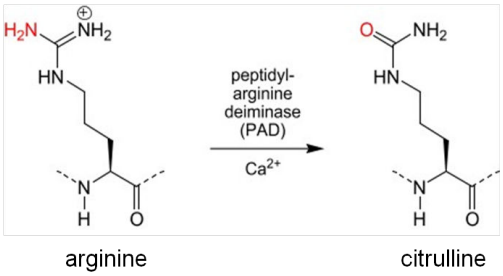


Figure 15 : réaction de conversion d'une arginine en citrulline par une enzyme de la famille des PAD. D'après ChemgaPedia (http://www.chemgapedia.de).

L'ubiquitinylation et la sumoylation correspondent respectivement au transfert d'une unité ubiquitine (Ub) de 8,5 kDa et d'une unité SUMO (Small Ubiquitin-like Modifier) de 12 kDa. Ces deux entités protéiques possèdent une homologie de séquence d'environ 20% et présentent une structure tridimensionnelle très similaire (figure 16). L'ubiquitinylation est catalysée par des enzymes de la famille des

lysine-ubiquitinases et la sumoylation par des enzymes de la famille des SAE (*SUMO-activating enzymes*). A l'image de la méthylation, ces deux modifications peuvent aléatoirement être des marques d'activation ou de répression de la transcription. Leur effet activateur de la transcription est du à l'encombrement stérique que ces deux gros polypeptides génèrent ce qui stabilise la chromatine dans une conformation ouverte, tandis que leur effet répresseur s'explique en partie par le masquage de certains résidus cibles d'autres modifications post-traductionnelles qui peuvent être activatrices de la transcription. Cependant, la sumoylation semble être davantage associée à une répression de la transcription en favorisant le recrutement d'histone-désacétylases[77]. L'ubiquitinylation est pour sa part considérée comme une marque répressive de la transcription lorsqu'elle affecte H2A et comme une marque alternativement activatrice ou répressive lorsqu'elle affecte H2B[78]. Par ailleurs, les histones ne peuvent être que mono-ubiquitinylées et non poly-ubiquitinylées comme c'est le cas lors de l'adressage des protéines au protéasome.

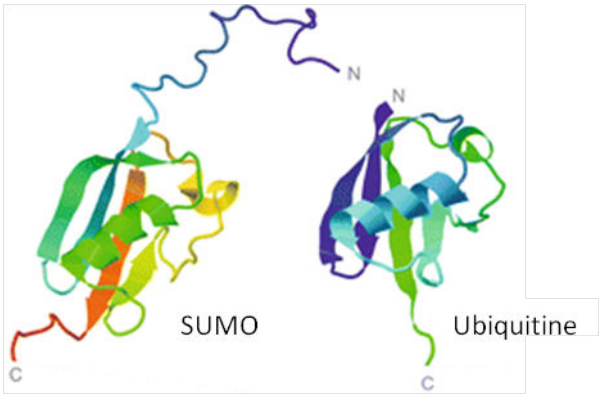


Figure 16 : structure des groupements SUMO et Ubiquitine chez l'Homme. D'après[79].

La ribosylation est une modification post-traductionnelle assez courante qui consiste au transfert d'un groupement poly (ADP-ribose) (figure 17) sur un résidu lysine depuis une molécule de NAD^+ par l'intermédiaire d'une enzyme de la famille des PARP (poly (ADP-ribose) polymérase). Dans le cas des histones, la ribosylation affecte la charge portée par les lysines et entraine une perturbation des interactions électrostatiques entre l'ADN et les histones, aboutissant à un relâchement de la chromatine. Elle joue également un rôle important dans les mécanismes de réparation des cassures de l'ADN[80].

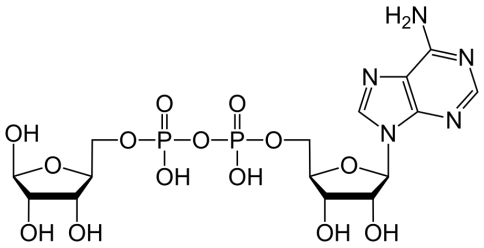


Figure 17 : structure du groupement ADP ribose d'environ 560 Da transféré sur une lysine par une enzyme de la famille des PARP lors d'une ribosylation.
(http://commons.wikimedia.org/wiki/File:Adenosine_diphosphate_ribose.svg#mediaviewer/File:Adenosine_diphosphate_ribose.svg)

La biotinylation est elle aussi une modification post-traductionnelle qui affecte exclusivement les résidus lysines. Elle correspond au transfert d'un groupement biotine d'environ 244 Da (figure 18) catalysé par des enzymes de la famille des biotinidases. Du point de vue biologique, la biotinylation des histones semble être associée à l'hétérochromatine[81] et donc impliquée dans la répression de la transcription des gènes. Elle joue également un rôle dans la signalisation des dommages à l'ADN[82].

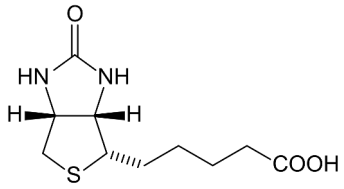


Figure 18 : structure du groupement biotine également connu sous le nom de Vitamine B8. (http://commons.wikimedia.org/wiki/File:Biotin_structure.svg#mediaviewer/File:Biotin_structure.svg).

Enfin, la crotonylation, découverte en 2011 par Tan M. *et al.*[83], est la modification la plus récemment mise en évidence sur les résidus lysines des histones. Le groupement crotonyle est transféré depuis une molécule de crotonyl-CoA, mais les enzymes qui catalysent cette réaction restent inconnues à ce jour (figure 19). De par sa structure proche d'une acétylation, la crotonylation aura la même répercussion sur la conformation de la chromatine. Elle est donc associée à une augmentation de l'activité transcriptionnelle des gènes.

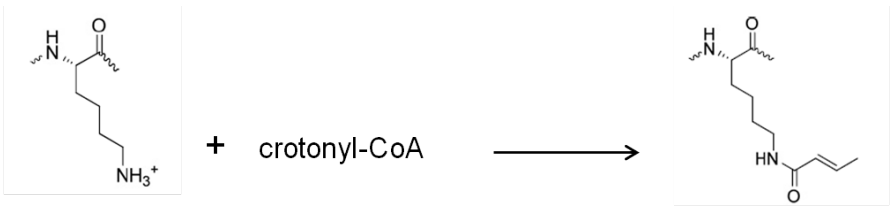


Figure 19 : réaction de transfert d'un groupement crotonyle sur un résidu lysine depuis une molécule de crotonyl-CoA.

II.6 Les partenaires d'interaction des histones modifiées

Nous venons de décrire comment les modifications post-traductionnelles des histones agissent sur la conformation de la chromatine et donc sur l'expression des gènes. Mais ce n'est pas leur seule façon de réguler l'activité transcriptionnelle. Ces modifications peuvent recruter divers partenaires protéiques et être à l'origine d'interactions protéines-protéines spécifiques impliquant de véritables lecteurs épigénétiques. Cependant, il semble qu'un seul résidu modifié ne soit pas suffisant pour recruter un complexe de remodelage, démontrant ainsi davantage l'importance primordiale de l'aspect combinatoire du code histone. Nous nous intéresserons dans ce chapitre aux domaines d'interactions avec des résidus acétylés, méthylés ou phosphorylés.

II.6.1 Domaines d'interaction avec les résidus acétylés

Le bromodomaine est un domaine protéique composé de quatre hélices α (Z, A, B et C) reliées par des boucles qui constituent une poche hydrophobe capable de se lier spécifiquement aux résidus lysine acétylés (figure 20). Seul domaine d'interaction avec la chromatine capable de reconnaitre les résidus acétylés, il a été découvert pour la première fois chez la drosophile au sein de la protéine Brahma[84] et a ensuite été identifié dans plus d'une centaine d'autres protéines chez divers organismes. La fonction première du bromodomaine est de réguler l'expression des gènes en reconnaissant spécifiquement les résidus acétylés des histones. Chez l'Homme, il est présent dans la plupart des complexes à activité HAT, comme par exemple chez les protéines de la famille GNAT, celles de la famille CBP/p300 ou encore chez le complexe TFIID au niveau de la sous-unité

TAF1 (figure 20). Il est également présent au sein de certains complexes de remodelage de la chromatine ATP-dépendants tel que le complexe SWI/SNF[85].

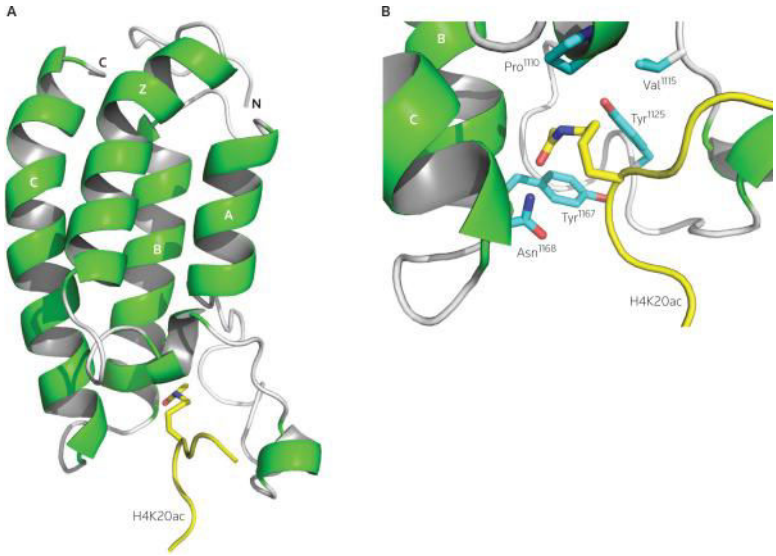


Figure 20 : A) structure 3D du bromodomaine de la protéine CBP se liant à une histone acétylée sur sa lysine 20 (H4K20ac). B) vue détaillée du domaine de liaison entre le bromodomaine et l'histone acétylée H4K20ac montrant les principales interactions entre acides aminés. D'après[86].

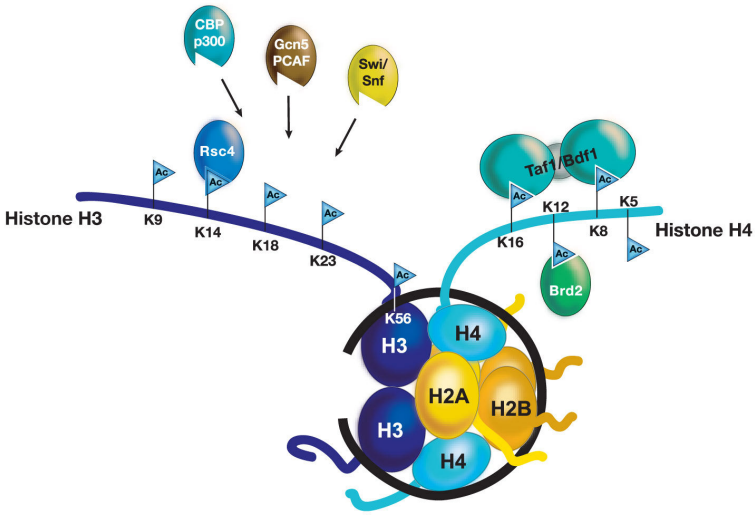


Figure 21 : schéma représentant différents sites d'acétylation des histones de cœur et quelques exemples de complexes qui possèdent un bromodomaine de liaison spécifique de ces résidus acétylés. Les acétylations sont représentées par des drapeaux bleus marqués Ac. D'après[6].

II.6.2 Domaines d'interaction avec les résidus méthylés

Contrairement au bromodomaine qui est le seul domaine d'interaction avec les lysines acétylées a avoir été identifié, il existe cinq domaines capables d'interagir spécifiquement avec les lysines méthylées[87]. Cette diversité des domaines de reconnaissance accroît considérablement le nombre de protéines capables de se lier aux résidus méthylés, et ce lors de nombreux processus cellulaires (figure 22).

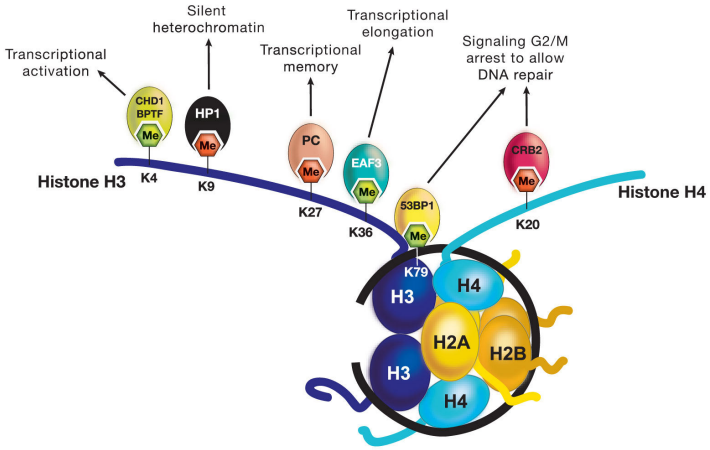


Figure 22 : schéma représentant différentes lysines méthylées des histones de cœur et quelques exemples de complexes de liaison spécifiques de ces résidus impliqués dans divers processus cellulaires. Les méthylations sont représentées par des hexagones (Me) verts lorsqu'il s'agit d'une marque d'activation et rouges lorsqu'il s'agit d'une marque de répression de la transcription. D'après[6].

II.6.2.1 Le chromodomaine

Le chromodomaine (*Chromatin Organization Modifier domain*) est un domaine structural composé d'un feuillet ß antiparallèle faisant face à une hélice α (figure 23). Egalement décrit pour la première fois chez la Drosophile dans certains complexes de remodelage de la chromatine dont HP1 (*Heterochromatin Protein 1*) et Polycomb[88], le chromodomaine assure une interaction spécifique avec les lysines méthylées, particulièrement sur l'histone H3. La plupart des protéines recrutées par les résidus méthylés *via* leur chromodomaine sont impliquées dans la formation d'hétérochromatine et la répression de la transcription[89].

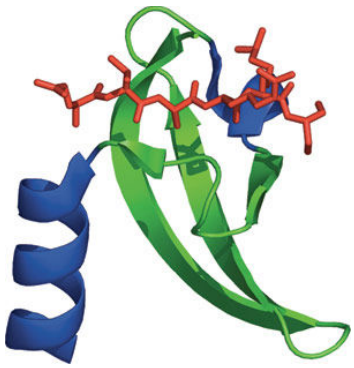


Figure 23 : représentation d'une structure cristallographique de l'interaction entre le chromodomaine de HP1 (en vert et bleu) et la lysine K9 méthylée de l'histone H3 (en rouge). D'après[90].

II.6.2.2 Le motif WD40

Le motif WD40, également appelé répétition ß-transducine, est un motif structural composé de quatre brins ß dont la séquence se termine par un dipeptide tryptophane-acide aspartique (W-D). La répétition de plusieurs motifs WD40 forme après repliement un domaine structural en turbine noté WD (figure 24). Les domaines WD peuvent contenir entre quatre et seize motifs WD40. Ils confèrent aux protéines qui le contiennent une spécificité vis-à-vis de certains résidus méthylés.

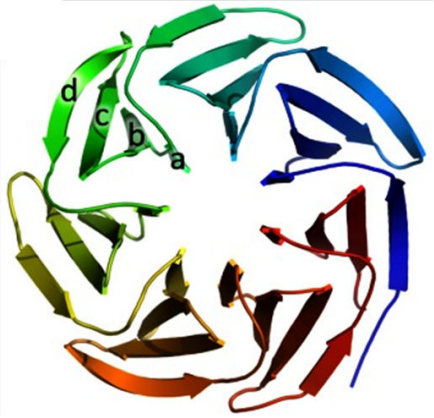


Figure 24 : structure d'un domaine WD composé de sept motifs WD40. D'après[91].

II.6.2.3 Le domaine Tudor

Le domaine Tudor est composé de cinq brins ß qui forment un feuillet antiparallèle présentant une architecture en tonneau (figure 25). Il a d'abord été identifié chez la Drosophile avant d'être également découvert chez l'Homme. Ce domaine a la particularité de reconnaître deux types de résidus méthylés. Ainsi, il s'agit du seul domaine connu capable de reconnaître les arginines diméthylées. Entre autres, il permet à la protéine JMJD2A à activité histone-déméthylase de se fixer aux résidus H3K4 triméthylés et H4K20 di- ou triméthylés.

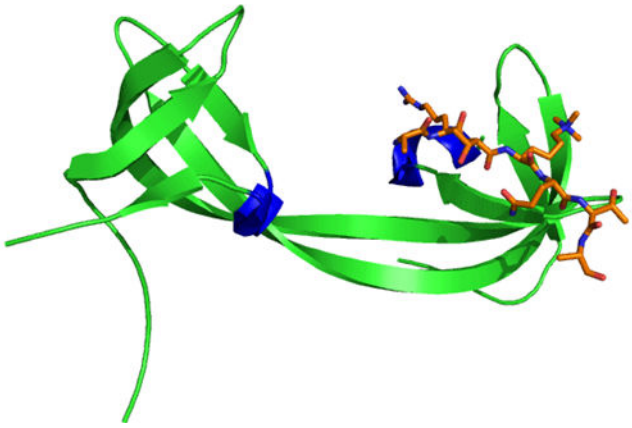


Figure 25 : structure 3D d'un domaine Tudor et interaction avec le résidu H3K4 triméthylé. D'après[92].

II.6.2.4 Le motif MBT

Le motif MBT (Malignant Brain Tumor) est composé d'une centaine d'acides aminés. Les protéines qui le contiennent sont capables de reconnaître certaines lysines méthylées, notamment sur H3 et H4, et sont majoritairement impliquées dans la répression de la transcription. En réalité, c'est la répétition de plusieurs motifs MBT formant un domaine structural globulaire hydrophobe qui permet l'interaction avec certains résidus spécifiques. C'est le cas de la protéine L(3)MBTL (figure 26) de la famille Polycomb qui est constituée de trois motifs MBT et qui est capable d'interagir spécifiquement avec le résidu H3K4 méthylé[93].

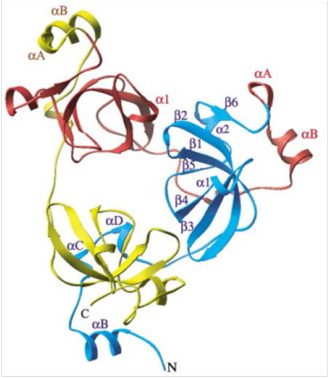


Figure 26 : structure 3D de la protéine L(3)MBTL composée de trois motifs MBT. D'après[93].

II.6.2.5 Le doigt de zinc de type PHD

Le doigt de zinc de type PHD (*Plant Homeo Domain*) est un motif structural composé d'environ soixante-dix acides aminés organisés en deux brins ß et une hélice α (figure 27). Découvert pour la première fois chez *Arabidopsis thaliana*[94], il a par la suite été retrouvé chez l'Homme dans plus de cent protéines dont la plupart interagissent avec la chromatine. Ce motif est présent au sein de certains activateurs transcriptionnels de la famille CBP/p300 ainsi que dans certains complexes à activité HMT, HDM ou encore HDAC. Il présente une affinité particulière pour les résidus lysines triméthylés.

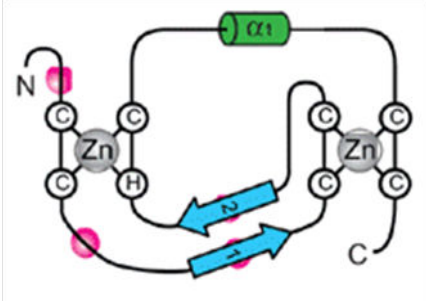


Figure 27: représentation topologique d'un domaine PHD. D'après[95].

II.6.3 Domaines d'interaction avec les résidus phosphorylés

Il n'existe actuellement que peu de protéines identifiées comme étant capables d'interagir spécifiquement avec des résidus phosphorylés d'histones. Seuls deux domaines structuraux, SH2 (*Src Homology region 2*) et BRCT (*BReast Cancer-susceptibility protein-1 C-Terminal)* ont été caractérisés. Le domaine SH2 est

présent dans plusieurs protéines à activité kinases et phosphatases. Quant au domaine BRCT, il est présent au sein de protéines impliquées dans la réparation de l'ADN et dans la régulation du cycle cellulaire.

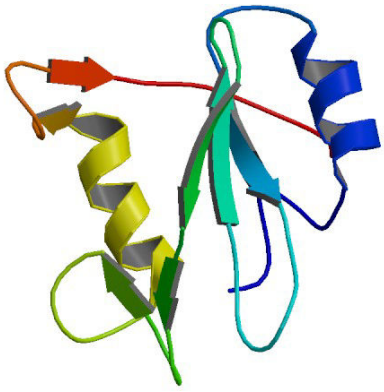


Figure 28 : structure 3D du domaine SH2 composé d'un feuillet ß antiparallèle et de deux hélices α. D'après[96].

III. Épigénétique, environnement et toxicologie

Comme nous venons de le voir au chapitre précédent, la régulation épigénétique est contrôlée par plusieurs mécanismes moléculaires qui sont interconnectés. Ces mécanismes sont extrêmement dynamiques et créent ensemble un équilibre nécessaire à l'expression normale des gènes tout au long de la vie d'une cellule. L'information épigénétique, ou épigénome, qui en découle est donc extrêmement complexe, tant qualitativement que quantitativement. Le meilleur exemple en est sans doute le code histone que nous venons de décrire. Sa complexité est issue tant de son aspect combinatoire que de la diversité des lecteurs épigénétiques qu'il peut recruter. Il présente ainsi deux niveaux de lecture différents. Certaines modifications post-traductionnelles sont stables au fil des divisions cellulaires et constituent une trame de fond qualifiée de mémoire épigénétique. C'est cette mémoire qui sera transmise de génération en génération. Sur cette trame héréditaire viennent s'ajouter des modifications dynamiques qui permettent des changements d'état transcriptionnel rapides et réversibles en réponse à des stimuli exogènes ou endogènes. Ce deuxième niveau de lecture s'apparente à la transduction d'un signal et peut impliquer des modifications de même nature que celles transmises de génération en génération. Autrement dit, au-delà de sa régulation physiologique au cours des étapes du développement embryonnaire ou du cycle cellulaire, la nature de l'épigénome peut être influencée par des facteurs environnementaux. La méthylation et l'hydroxyméthylation de l'ADN, le code histone et les ARNmi peuvent être concernés par cette perturbation environnementale. Nous nous intéresserons tout au long de ce chapitre à l'impact de l'environnement et plus particulièrement aux expositions à des toxiques durant les phases précoces du développement.

De très nombreuses études s'intéressent aux effets génotoxiques et mutagènes des xénobiotiques environnementaux. Nous savons ainsi que de nombreuses substances chimiques sont capables d'induire des cassures de la molécule d'ADN, ou encore d'y former des adduits covalents[97-99]. Le lien de cause à effet entre une exposition à ces molécules génotoxiques et la survenue de différentes pathologies est donc sans ambiguïté. Ces maladies à composante environnementale sont très nombreuses : cancers, diabète, maladies métaboliques, maladies auto-immunes,

maladies neurodégénératives. Selon certains, le développement d'une maladie sur quatre pourrait être imputé à des facteurs environnementaux[100]. Il serait donc logique de s'intéresser de la même manière à l'ensemble des mécanismes cellulaires susceptibles d'y être sensible. Pourtant, malgré leur capacité à contrôler l'expression des gènes, les acteurs de la régulation épigénétique ne font que rarement partie des mécanismes incriminés. L'impact épigénétique d'une exposition aux xénobiotiques reste de nos jours trop souvent sous-estimé et peu étudié. Nous tenterons donc au cours de ce chapitre de rappeler combien il est primordial de s'intéresser aux effets épigénétiques d'une exposition aux molécules chimiques à laquelle nous sommes quotidiennement soumis à travers l'environnement. Nous insisterons plus particulièrement sur les risques pour la santé d'une perturbation épigénétique lors de périodes de vulnérabilité telles que la grossesse et le développement embryonnaire. Comme c'est le cas pour le reste du manuscrit, la discussion se fera sous l'angle des histones et de leurs modifications post-traductionnelles.

III.1 Grossesse et pathologies

III.1.1 Les hypothèses de Barker

Le concept d'origine fœtale des maladies a émergé il y a plus de 25 ans à la suite d'études épidémiologiques sur la mortalité infantile et adulte. Une série de trois articles publiés entre 1986 et 1993 par Barker *et al.*[101-103] dans la célèbre revue britannique *The Lancet* a posé les bases fondatrices de l'hypothèse de l'origine fœtale des maladies, connue également sous le nom d'hypothèse de Barker. Les premiers travaux de Barker ont ainsi montré à partir de données épidémiologiques qu'il existait une corrélation entre le taux de mortalité infantile entre 1921 et 1925 et le taux de maladies cardiaques ischémiques entre 1968 et 1978 en Angleterre et au Pays de Galles. Barker a expliqué cette corrélation par plusieurs facteurs sociaux et environnementaux. Ces observations l'ont conduit à conclure que la corrélation entre les taux de mortalité infantile et adulte était géographie dépendante. Il émit l'hypothèse que des variations nutritionnelles subies en début de vie s'exprimaient de façon pathologique sous l'influence d'expositions alimentaires ultérieures, introduisant ainsi la notion d'effet retard.

Quelques années plus tard, Barker poussa ses investigations plus loin en utilisant une nouvelle cohorte d'adultes. En corrélant la taille et le poids de nouveaux-nés au taux de mortalité adulte à la suite d'une maladie cardiaque ischémique, il a émis une deuxième hypothèse affirmant qu'un environnement fœtal donnant lieu à un retard de croissance du fœtus et de l'enfant avait pour conséquence une susceptibilité accrue aux maladies cardiaques ischémiques à l'âge adulte.

Enfin, Barker et ses collaborateurs ont fini par démontrer comment une dénutrition fœtale à différents stades de la gestation pouvait être liée à des phénotypes anormaux à la naissance. Ils ont montré que chacun de ces phénotypes était issu d'adaptations physiologiques se traduisant par des variations de concentrations de certaines hormones placentaires, entrainant par la suite différents désordres métaboliques à l'âge adulte.

Tous ces travaux ont conduit la communauté scientifique à prendre conscience de l'impact majeur de l'environnement fœtal sur la santé et le développement de maladies de l'adulte, proposant ainsi le concept de plasticité développementale[104] dont le placenta serait l'acteur principal.

III.1.2 Le placenta : organe clé dans la plasticité développementale

Le rôle clé du placenta dans l'origine développementale des maladies a clairement été démontré par les travaux de Barker et ses collaborateurs. Il convient de rappeler quelques éléments de physiologie placentaire qui peuvent permettre de mieux comprendre comment cet organe réagit aux différents stimuli auxquels il est soumis au cours de la grossesse, et quelles seront les conséquences de ces adaptations pour le fœtus.

III.1.2.1 Rappels physiologiques

Le placenta est un organe transitoire assurant le maintien et le bon déroulement de la grossesse. Sa structure est extrêmement hétérogène à travers les espèces. Chez l'Homme, il présente une physiologie très complexe et assure pour le fœtus à la fois le transport de l'oxygène, les apports énergétiques et

l'élimination des déchets. C'est une véritable interface entre la mère et le fœtus que l'on pourrait qualifier de barrière materno-fœtale, qui possède également des fonctions endocrines indispensables au maintien de la gestation. A son terme, le placenta présente une structure ovale d'environ 20 cm pesant jusqu'à 500g[105]. L'unité structurale et fonctionnelle du placenta est appelée villosité choriale (figure 29). Elle est constituée de plusieurs composants de nature différente : les cytotrophoblastes villeux (CT), le syncytiotrophoblaste, le mésenchyme, les cellules de Hofbauer et les vaisseaux fœtaux.

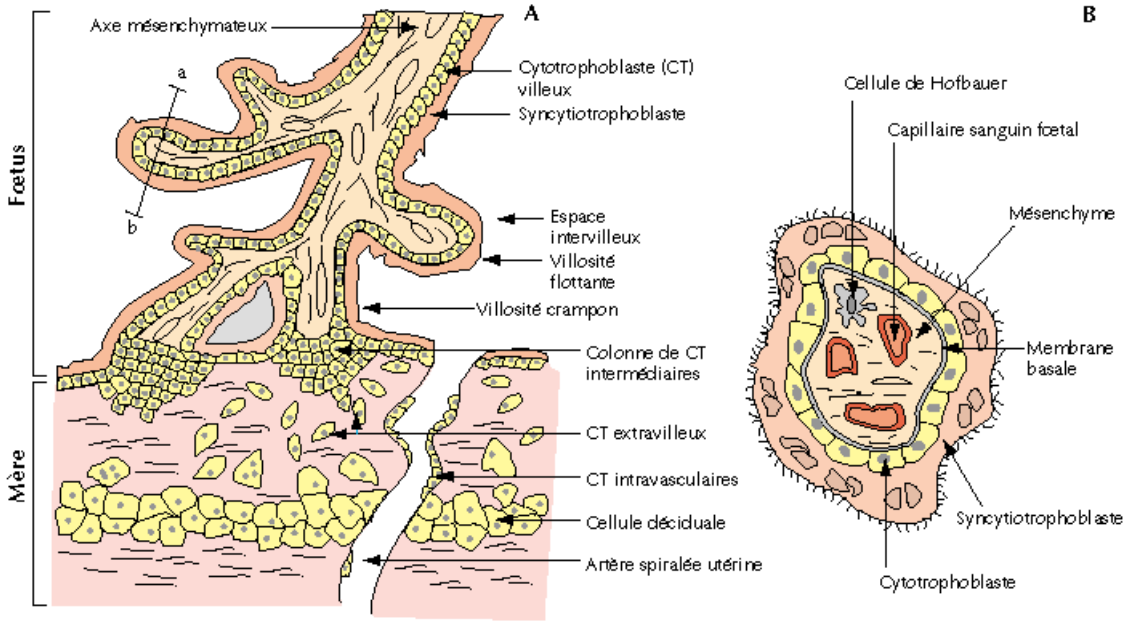


Figure 29 : schéma structural du placenta humain représentant en A) une coupe longitudinale et en B) une coupe transversale d'une villosité choriale. D'après[106].

Le principal type cellulaire du placenta est donc le cytotrophoblaste qui peut se différencier en cytotrophoblastes villeux ou extravilleux qui auront chacun des fonctions différentes (figure 30). En fusionnant, les cytotrophoblastes villeux donneront naissance au syncytiotrophoblaste qui est le tissu fonctionnel constitutif du placenta, responsable à la fois des fonctions endocrines et d'échange. C'est un véritable tissu sexué qui possède le même caryotype que le fœtus.

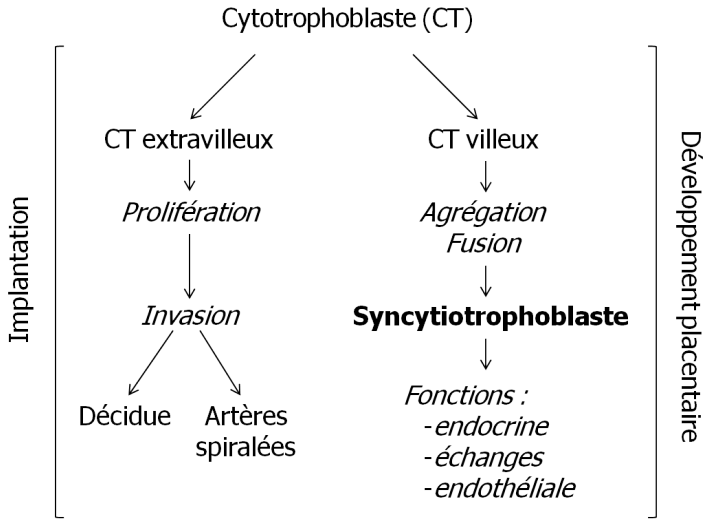


Figure 30 : voies de différenciation du cytotrophoblaste et fonctions associées. D'après[105].

Le réseau vasculaire du placenta est un système clos. L'intégrité du syncytiotrophoblaste garantit l'absence de contact entre le sang fœtal et le sang maternel présent dans la chambre intervilleuse. Malgré cela, cette barrière est loin d'être imperméable. Elle laisse passer tous les micro- et macronutriments dont le fœtus a besoin, mais également les virus, les bactéries et certains xénobiotiques. Les mécanismes d'échange entre la mère et le fœtus évoluent au cours de la grossesse, et on en distingue deux types : la diffusion passive et le transport actif réalisé par des protéines membranaires. Ainsi les gaz, l'eau, les ions, les acides aminés, les lipides ou encore les immunoglobulines franchissent la barrière syncytiotrophoblastique par diffusion ou par transport actif.

Du côté endocrine, le placenta synthétise plusieurs hormones polypeptidiques et stéroïdes nécessaires à la croissance du fœtus et au bon déroulement de la grossesse. Le syncytiotrophoblaste synthétise ainsi trois hormones polypeptidiques majeures : l'hormone gonadotrophique chorionique humaine (hCG), l'hormone lactogène placentaire (hPL) et l'hormone de croissance placentaire (GH). Il synthétise également certaines hormones stéroïdiennes telles que la progestérone ou les œstrogènes. Chacune de ces hormones a des fonctions bien précises qui sont temps- et concentration-dépendantes au cours de la gestation. Ces hormones placentaires assurent la qualité de la placentation au cours du premier trimestre de grossesse en assurant le remaniement vasculaire utérin. Dès l'arrivée du sang maternel dans la chambre intervilleuse vers la fin du premier trimestre, ces

hormones permettent à l'organisme maternel de s'adapter au maintien de la grossesse et de répondre aux besoins énergétiques du fœtus.

Enfin, le placenta a également des capacités de métabolisation des xénobiotiques *via* la présence de nombreuses enzymes de phase I, notamment les cytochromes P450 (CYP). Plusieurs isoformes de CYP sont exprimées dans les mitochondries et le réticulum endoplasmique des cellules trophoblastiques[107]. Dans le placenta humain à terme, les isoformes CYP1A1, 2E1, 3A4, 3A5, 3A7 et 4B1 ont été caractérisées à l'échelle protéique, les autres n'ayant été détectées qu'à l'échelle des transcrits (ARNm)[108,109]. Généralement, l'expression des CYP varie au cours de la grossesse et est maximale au cours de premier trimestre de grossesse, période durant laquelle le fœtus est le plus vulnérable.

III.1.2.2 Adaptation du placenta à l'environnement materno-fœtal et conséquences sur la programmation fœtale.

La physiologie et les fonctions du placenta montrent combien c'est un organe clé durant la gestation. Son intégrité est garante du bon déroulement de la grossesse et de la viabilité du fœtus. Paradoxalement, c'est un organe qui fait preuve d'une remarquable plasticité et dont la structure ainsi que les fonctions s'adaptent aux besoins du fœtus qui évoluent tout au long de la gestation. On distingue ainsi des adaptations morphologiques (vascularisation, épaisseur de la barrière hémato-placentaire, composition cellulaire) et des adaptations fonctionnelles (transport des micro- et macronutriments). Malheureusement, cette plasticité rend le placenta vulnérable à l'ensemble des facteurs environnementaux auxquels la mère sera exposée au cours de la grossesse. Parmi eux, de nombreuses substances toxiques : médicaments, alcool, tabac et autres xénobiotiques environnementaux.

Le tabagisme nous intéressera plus particulièrement au cours d'un prochain chapitre. Qu'il soit actif ou passif, il est reconnu comme étant le principal facteur de risque environnemental des complications de la grossesse. L'exposition de la femme enceinte, quel que soit le stade de la grossesse, peut être associée à un retard de croissance intra-utérin, une mort intra-utérine, une insuffisance placentaire ou encore une naissance prématurée[106]. Ceci s'explique par le fait que

la fumée de cigarettes contient de très nombreux composés chimiques toxiques (nicotine, cyanures, sulfures, hydrocarbures aromatiques, cadmium) qui traversent facilement et rapidement la barrière placentaire. On observe ainsi un remodelage anatomique du placenta chez les femmes enceintes exposées au tabac, ainsi qu'une modulation de ses activités hormonales et enzymatiques. En effet l'expression des CYP du placenta peut être modulée par le tabac ou par d'autres contaminants environnementaux présents dans l'air ou l'alimentation[110]. Cette modulation rendra le placenta encore plus sensible vis-à-vis de certains toxiques. De nombreuses autres substances génotoxiques ou considérées comme des perturbateurs endocriniens auront également des effets délétères sur le placenta et finalement sur le fœtus lui-même, point que nous ne détaillerons pas ici.

Face à ces agressions quotidiennes, le placenta peut modifier l'environnement du fœtus tant du point de vue hormonal que de la délivrance de certains gaz ou nutriments (figure 31).

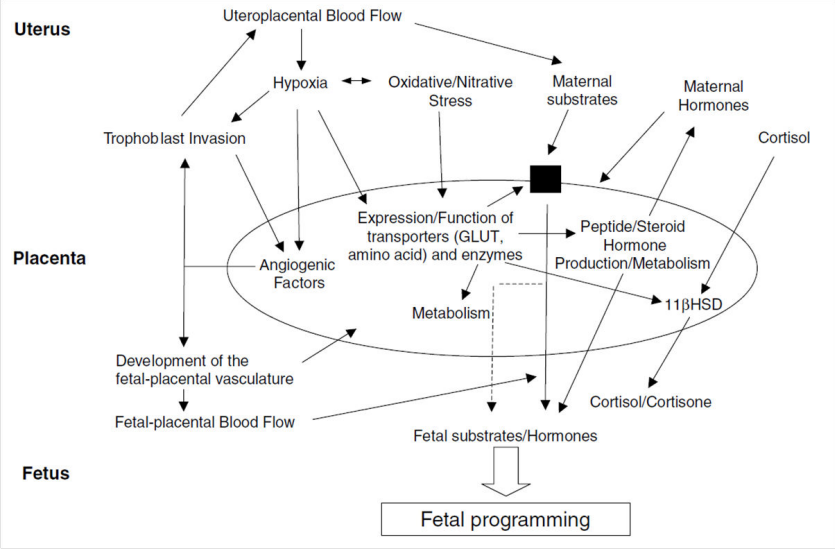


Figure 31 : réponses adaptatives du placenta et conséquences sur la programmation fœtale. D'après[111].

Plusieurs mécanismes moléculaires peuvent donc être impliqués dans la plasticité placentaire, mais ils restent cependant très peu connus à cause de la complexité des interactions entre la mère, le placenta et le fœtus. Certaines études récentes suggèrent qu'il existerait deux principaux mécanismes capables d'assurer cette plasticité : les voies de détection des nutriments et les mécanismes épigénétiques[112]. Deux principales voies de détection des nutriments semblent être

impliquées dans la plasticité placentaire. Il s'agit de la voie TOR (*target of rapamycin*) et de la voie AAR (*amino acid response*) qui permettent au placenta de se comporter comme un détecteur de nutriments et de réguler directement le transport de certains composés endogènes. En parallèle, les mécanismes épigénétiques semblent être également capables d'intégrer les signaux environnementaux au niveau placentaire et de les répercuter à l'échelle de la chromatine en modulant l'expression de certains gènes.

III.1.2.3 Mécanismes épigénétiques et plasticité placentaire

Les mécanismes épigénétiques sont les principaux acteurs de l'interaction entre les gènes et l'environnement, notamment à l'échelle placentaire. Ils définissent les profils d'expression fœtale des gènes et pourront modifier de façon permanente ou transitoire leur expression sous l'influence de facteurs environnementaux. Ces mécanismes de régulation sont notamment impliqués dans les phénomènes d'empreinte parentale des gènes. Chez l'Homme, le niveau d'expression de l'allèle paternel et de l'allèle maternel est comparable pour la plupart des gènes autosomiques[113]. Cependant, les gènes soumis à empreinte s'expriment de manière mono-allélique dans les cellules somatiques et les tissus. La copie muette peut être l'allèle maternel ou paternel, et le choix se fait de manière totalement aléatoire durant le développement embryonnaire[114]. Les trois mécanismes épigénétiques décrits précédemment seraient impliqués dans le « silençage » d'un des deux allèles. Or dans le placenta, plusieurs gènes sont soumis à empreinte. Certaines études suggèrent que les phénomènes d'empreinte au niveau placentaire impliqueraient majoritairement des modifications post-traductionnelles d'histones dont la méthylation des lysines sur l'histone H3, ainsi que des ARNmi[113]. Mais il n'y a pas que les gènes soumis à empreinte qui sont sujets à régulation épigénétique au niveau du placenta. Plus généralement, la régulation épigénétique de l'expression fœtale des gènes est fondamentale pour le développement placentaire et la programmation fœtale. Toute modulation indésirable par l'intermédiaire des mécanismes épigénétiques pourra être très lourde de conséquences pour l'enfant en devenir, mais également pour la mère elle-même (figure 32). Plusieurs pathologies de la grossesse sont ainsi associées à

des perturbations épigénétiques : retard de croissance intra-utérin, pré-éclampsie, voire même une mort in *utero*.

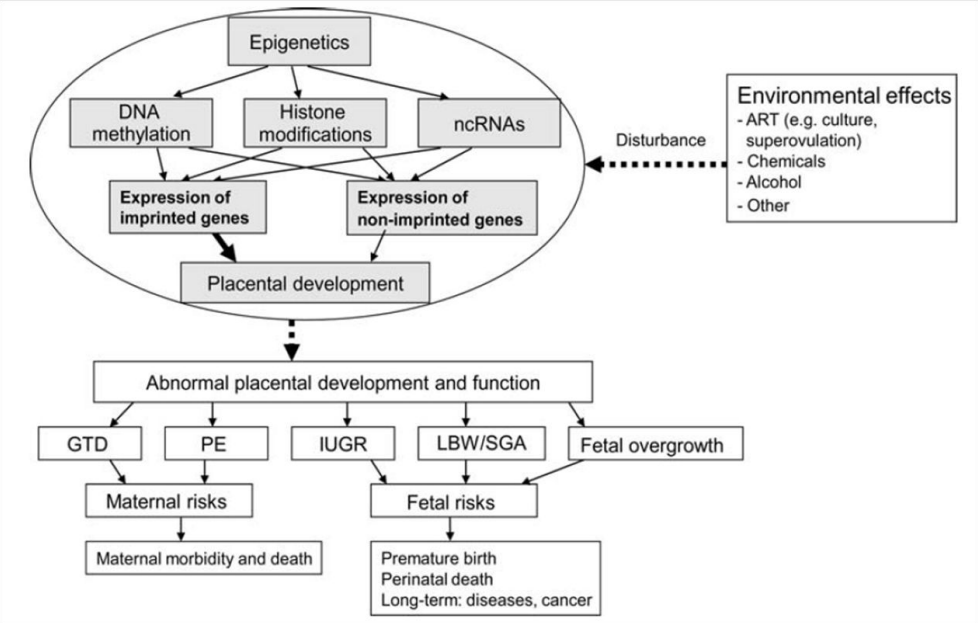


Figure 32 : schéma récapitulatif des mécanismes épigénétiques impliqués dans la plasticité placentaire et conséquences de leur perturbation sur le placenta. ncRNAs = ARN non codants, ART = techniques de procréation médicalement assistée, GTD = maladie trophoblastique gestationnelle, PE = pré-éclampsie, IUGR = retard de croissance intra-utérin, LBW = faible poids à la naissance, SGA = petit pour l'âge gestationnel. D'après[115].

III.1.3 Les origines développementales de la santé et des maladies de l'adulte : perturbation environnementale de l'épigénome

A la lumière des hypothèses de Barker, l'étude des phénomènes de plasticité développementale et placentaire médiés par des mécanismes épigénétiques permet d'aboutir au concept d'origine développementale de la santé et des maladies de l'adulte (DOHaD). Les mécanismes épigénétiques font partie des moyens dont dispose le placenta et plus particulièrement la cellule trophoblastique pour s'adapter à l'environnement maternel. En plus d'induire des pathologies de la grossesse, l'exposition placentaire aux toxiques aura pour conséquence une prédisposition à certaines maladies à l'âge adulte. Elle pourra avoir des effets retardés héréditaires, parfois sur plusieurs générations sans même que l'exposition ne soit maintenue[116]. L'épigénome représente ainsi la base moléculaire de la susceptibilité à développer certaines maladies. Il faut tenter de comprendre par quels mécanismes l'environnement perturbe l'épigénome, puis les conséquences

que cela aura sur la viabilité du fœtus et du futur adulte. Moshe Szyf dans sa revue sur l'implication de l'épigénome en toxicologie[117] présente plusieurs mécanismes par lesquels les toxiques perturbent l'épigénome. D'après lui, les toxiques peuvent interagir directement avec les enzymes qui régulent la méthylation de l'ADN (DNMT, déméthylases) ou les modifications post-traductionnelles des histones (HAT, HDAC, HMT, HDM). Ils peuvent également les priver de leurs substrats tels que la SAM ou l'acétyl-CoA. Quel que soit le mécanisme de perturbation, cela aboutira à une perturbation des profils de méthylation de l'ADN et à l'apparition de marques épigénétiques aberrantes sur les histones.

Il est donc fondamental pour l'évaluation des risques de rechercher et d'identifier les potentiels effets épigénotoxiques des xénobiotiques auxquels la femme enceinte est exposée de façon aigüe ou chronique. Cette exploration permettra de mieux comprendre et de prédire les effets délétères de ces toxiques à court, moyen et long terme.

III.2 Code histone et exposition aux toxiques

Ce manuscrit s'intéressant particulièrement aux histones et à leurs modifications post-traductionnelles, nous allons passer en revue l'ensemble des données disponibles à ce jour dans la littérature attestant d'une perturbation du code histone après une exposition volontaire ou involontaire à divers xénobiotiques. Il faut cependant garder en mémoire que la plupart de ces perturbations du code histone s'accompagnent d'une perturbation des profils de méthylation de l'ADN, et parfois de l'expression des ARNmi. Les études que nous allons passer en revue permettront de mieux comprendre pourquoi il est important de rechercher des marqueurs histoniques d'une perturbation de l'épigénome en cas d'exposition à des toxiques.

Le code histone, de par sa nature dynamique, est une cible de choix pour les toxiques. Les preuves s'accumulent en faveur du rôle que joue la perturbation du code histone dans le développement de nombreuses pathologies. Lorsque l'exposition à des toxiques se fait in *utero*, nous venons de voir quelles pouvaient en être les conséquences sur le développement futur de maladies chez l'adulte. Mais cette exposition peut naturellement se faire en dehors des périodes de

gestation, tout au long de la vie, et n'en sera pas moins lourde de conséquences. Chez l'adulte, des marques histoniques aberrantes sont souvent retrouvées dans différents types de cancers, mais également dans des maladies neurodégénératives. Toutes ces pathologies ont un point commun : une forte composante environnementale. L'exposition à des agents neurotoxiques tels que les métaux lourds ou les pesticides est considérée aujourd'hui comme un véritable facteur de risque dans le développement de maladies neurodégénératives chroniques, dont la maladie d'Alzheimer et la maladie de Parkinson[118] qui représentent actuellement un véritable problème de santé publique. Même constat du côté cancérologie, considéré comme le mal du XXI^{ème} siècle. La quasi-totalité des cancers chez l'Homme présente une composante épigénétique avec la présence de marques aberrantes. L'implication du code histone dans ces pathologies est telle que de nombreuses stratégies thérapeutiques utilisent des modulateurs épigénétiques ciblant spécifiquement les enzymes de modifications des histones.

Les perturbations épigénétiques ne sont pas toujours facilement imputables à une exposition environnementale à des toxiques ou à une pathologie particulière. En effet, ces changements sont proportionnellement très faibles, cumulatifs et temps-dépendants. La frontière entre modification épigénétique adaptative et modification épigénétique délétère est très mince, ce qui complique l'établissement clair de relations de cause à effet dans le cas du développement d'une pathologie chronique suite à une exposition à un toxique[119]. De plus, l'Homme est dans la majorité des cas exposé à travers l'environnement à un mélange de toxiques aboutissant à ce que l'on appelle l'effet « cocktail » des polluants[120]. Mais si l'on considère, comme l'affirmait Paracelse, que « toutes les choses sont poison, et rien n'est sans poison ; seule la dose détermine ce qui n'est pas un poison », tous les éléments avec lesquels nous sommes en contact, y compris notre alimentation, sont susceptibles d'induire des modifications de notre épigénome[121]. D'un point de vue toxicologique, de nombreuses études mécanistiques ont relié l'exposition *in vitro* ou *in vivo* à un toxique avec l'apparition de marques histoniques aberrantes. La plupart de ces études concernent les métaux lourds et les pesticides, et dans une moindre mesure les hydrocarbures aromatiques polycycliques.

III.2.1 Les métaux lourds

Le Dr. Broday et ses collaborateurs ont été parmi les premiers à étudier l'effet d'un toxique sur les histones et leurs modifications post-traductionnelles. En 2000, ils ont publié un article montrant les effets de l'exposition *in vitro* de cellules de mammifères au nickel (Ni) sur l'acétylation de l'histone H4[122]. Le Ni est un métal de transition naturellement présent dans divers minerais et utilisé en industrie pour la fabrication de batteries alcalines ou la production d'aciers inoxydables. Le Nickel et ses nombreux dérivés sont connus pour leurs propriétés carcinogènes. Il a ainsi été montré qu'une exposition au Ni à des doses subtoxiques (non carcinogènes) induisait une diminution globale de l'acétylation de H4, ciblant particulièrement la lysine 12. D'autres études sur les effets épigénétiques du Ni ont suivi. En 2003, l'équipe du Dr. Zhang a mis en évidence qu'un prétraitement de cellules humaines par la trichostatine A (TSA), inhibiteur des HDAC, diminuait significativement le processus de carcinogenèse après une exposition au Ni[123]. Par la suite, d'autres perturbations des modifications post-traductionnelles des histones induites par le Ni ont été mises en évidence : diminution globale de l'acétylation des histones H2A, H2B, H3 et H4, augmentation de la diméthylation de H3K9 et augmentation de l'ubiquitinylation de H2A et H2B[124]. Enfin, l'étude la plus récente publiée en 2012 par Arita *et al.* [125] et réalisée *in vivo* sur 45 sujets exposés au Ni a confirmé les résultats précédents et a également identifié certaines marques histoniques supplémentaires, telle qu'une augmentation de la triméthylation de H3K9.

L'arsenic (As) est un élément chimique semi-métallique naturellement présent dans la croûte terrestre et qui se retrouve à des concentrations élevées dans les sols et les eaux souterraines. Cet élément inorganique est également présent dans le tabac et est considéré comme carcinogène chez l'Homme[126]. Les effets sur les modifications des histones de deux de ses principaux dérivés oxydés, l'arsénite (As(III)) et l'acide monométhylarsonique (MMA(III)) sur les modifications d'histones ont été étudiés. La première étude faite sur des cellules humaines a été publiée en 2008 par Zhou *et al.*[127]. Elle démontre une augmentation de la di- et de la triméthylation de H3K4 après une exposition à l'As(III), associé à une diminution de la triméthylation de H3K27. L'équipe de Jensen et ses collaborateurs ont montré

que As(III) et MMA(III) pouvaient également perturber le degré de méthylation de l'histone H3 et induire l'expression aberrante de certains gènes impliqués dans les processus de carcinogénèse[128]. Cette même équipe a mis en évidence dans une seconde étude que l'histone H3 pouvait être hyperacétylée sous l'effet de As(III) et MMA(III), et que cette marque chromatinienne permissive pouvait se transmettre aux générations suivantes sans que l'exposition ne soit maintenue[129]. Mis à part H3, d'autres histones de cœur peuvent également être impactées par une exposition à l'As. C'est le cas de l'histone H4 dont l'acétylation sur la lysine 16 (H4K16) diminue après une exposition *in vitro* de cellules urothéliales humaines à As(III) ou MMA(III)[130].

Le cadmium est un métal de transition écotoxique présent dans les rejets industriels ainsi que dans la fumée de cigarette, certains engrais ou encore certains champignons qui peuvent accumuler des doses très élevées. L'exposition de cellules humaines au cadmium (Cd) semble avoir des effets comparables à une exposition à As(III) au niveau des modifications post-traductionnelles des histones. En effet, le Cd sous sa forme oxydée Cd(II) présente les mêmes effets sur l'expression des gènes que As(III) dont nous venons de décrire l'impact sur le code histone[131].

Le chrome (Cr) est un métal de transition dont la forme trivalente Cr(III) est très utilisée dans l'industrie chimique et métallurgique. Schnekenburger *et al*. ont montré qu'une exposition *in vitro* au Cr diminuait la phosphorylation de H3S10 et la triméthylation de H3K4, ainsi que le niveau global d'acétylation de H3 et H4[132]. Le Cr peut ainsi activer ou réprimer l'expression de certains gènes impliqués dans la carcinogénèse à travers ses effets épigénétiques. Il peut également s'opposer à l'induction du CYP1A1 médiée par une activation du récepteur AhR par des ligands exogènes[132]. En parallèle, une autre étude a montré que la di- et la triméthylation de H3K9 augmentait après une exposition *in vitro* de cellules humaines cancéreuses au Cr hexavalent (Cr(VI))[133]. Ces marques épigénétiques se retrouvaient encore augmentées plusieurs jours après l'arrêt de l'exposition, signifiant qu'elles se transmettent au fil des divisions cellulaires.

III.2.2 L'éthanol

L'éthanol est un alcool connu pour être une des plus anciennes drogues récréatives. Classé comme composé tératogène, sa toxicité est multiple. Il est également connu pour induire l'expression de certains gènes qui codent pour des enzymes du métabolisme des xénobiotiques. Il perturbe également de nombreuses voies métaboliques dont notamment celles de l'acide folique et de la méthionine, ayant ainsi des conséquences sur la méthylation de l'ADN[134]. Plus récemment, il a été découvert que l'éthanol perturbait également le code histone. Deux marques histoniques semblent principalement être augmentées par une exposition à l'éthanol chez l'Homme : H3K27 triméthylée et H3K9 acétylée[135]. D'autres marques histoniques spécifiques ont été identifiées chez d'autres espèces après une exposition aigüe à l'éthanol. Une diminution de H3K9 diméthylée et une augmentation de H3K4 diméthylée ont ainsi été observées sur des primo-cultures d'hépatocytes de rat[136].

III.2.3 Le cobalt

Le cobalt (Co) existe naturellement sous différentes formes. Le chlorure de cobalt(II) ($CoCl_2$) est un composé inorganique carcinogène et écotoxique retrouvé dans l'environnement car très utilisé dans diverses industries pour la fabrication d'encres, de peintures, de fertilisants agricoles ainsi que pour la production de vitamine B12. L'exposition *in vitro* de cellules de carcinome pulmonaire humain A549 au $CoCl_2$ a révélé que ce composé induisait une augmentation de la triméthylation de H3K4, de la di- et triméthylation de H3K9, de la triméthylation de H3K27 et H3K36 ainsi que de l'ubiquitinylation de H2A et H2B[137]. En parallèle, il induit une diminution de la méthylation de H3K4 et du niveau global d'acétylation de H4[137]. Les ions Co(III) peuvent également rentrer en compétition avec les ions Fe(III) au niveau des sites actifs de certaines enzymes dont les enzymes de modifications des histones et perturber leur activité.

III.2.4 Les drogues

La consommation de drogues addictives étant un véritable problème de santé publique dans notre société, elles représentent une part non négligeable des xénobiotiques auxquels l'Homme peut être exposé. Il existe des drogues dites dures et d'autres dites douces, la distinction se faisant selon l'addiction qu'elles engendrent. Dans le cas des drogues dures, la plus consommée est la cocaïne. C'est un alcaloïde extrait de la feuille de coca qui a de puissantes propriétés stimulantes du système nerveux central. La prise chronique de cocaïne perturbe la plasticité neuronale. Ce phénomène semble être lié à la perturbation de certains mécanismes épigénétiques, notamment l'acétylation des histones. L'activité des enzymes HDAC est ainsi perturbée lors de la prise de cette drogue au long terme, aboutissant à un déséquilibre de la balance HAT/HDAC et à une augmentation globale du niveau d'acétylation des histones[138]. La perturbation de l'activité HDAC serait même un facteur déclenchant le passage d'une utilisation ponctuelle récréative à une addiction réelle[138]. La plupart des études sur l'effet de la cocaïne au niveau neuronal ont été réalisées chez la souris. Maze *et al.* ont ainsi mis en évidence que la prise de cocaïne réduisait la triméthylation de H3K9 au niveau cérébral [139]. Une autre étude plus ancienne avait également montré que la cocaïne augmentait la phosphorylation de H3S10 et l'acétylation de H4K5 chez la souris[140].

Du côté des drogues douces, le cannabis arrive en tête des consommations. Le principe actif aux propriétés psychotropes retrouvé dans le cannabis est le Δ-9-tétrahydrocannabinol (THC). Une étude très récente a montré que la consommation de THC pouvait causer des troubles immunologiques et altérer la réponse immunitaire cellulaire lymphocytes T-dépendante. Ce mécanisme toxique semble être en partie expliqué par une perturbation locale de certaines marques histoniques dont la triméthylation de H3K4, H3K27, H3K9 et H3K36 et également l'acétylation de H3K9. Ces perturbations sont à l'origine de l'augmentation et de la diminution parallèle de l'expression de certains gènes à l'origine de la modulation de la réponse immunitaire[141].

III.2.5 Les pesticides

Les pesticides représentent une large classe de polluants organiques persistants (POP). Reconnus pour leur toxicité, de très nombreuses études épidémiologiques et expérimentales prouvent qu'ils peuvent être à l'origine d'effets néfastes pour la santé humaine à la fois en cas d'intoxication aigüe et chronique. Ils peuvent être à l'origine de simples irritations cutanées et oculaires, mais également d'effets neurotoxiques, reprotoxiques voire cancérigènes[142]. L'équipe du Dr. Kanthasamy a publié une série d'articles entre 2010 et 2012 sur la neurotoxicité des pesticides médiée par des mécanismes épigénétiques[118,143,144]. Dans leur première étude sur le sujet parue en 2010, ils ont étudié les effets d'une exposition *in vitro* de primo-cultures de neurones dopaminergiques à la dieldrine, un insecticide organochloré. Ils ont ainsi mis en évidence une hyperacétylation temps-dépendante des histones de cœur H3 et H4[143]. Cette hyperacétylation a été attribuée à un dysfonctionnement du protéasome conduisant à l'accumulation d'enzymes à activité HAT. D'un point de vue toxicologique, ils ont montré que cette hyperacétylation de H3 et H4 aboutissait à une apoptose massive des neurones dopaminergiques, mécanisme de neurodégénérescence impliqué dans l'étiopathologie de la maladie de Parkinson. Ce mécanisme de toxicité a été confirmé en inhibant les HAT par l'acide anacardique, ce qui a conduit à une diminution significative de la dégénérescence des neurones dopaminergiques du mésencéphale. Ils ont ensuite étudié les effets neurotoxiques du paraquat, un des herbicides les plus utilisés au monde dont l'exposition chez l'Homme semble également être impliquée dans le développement de la maladie de Parkinson. En exposant une lignée cellulaire de neurones dopaminergiques au paraquat, ils ont mis en évidence une hyperacétylation de H3. Contrairement à la dieldrine, cette hyperacétylation ne semble pas affecter l'histone H4, et elle semble cette fois être médiée par une diminution de l'activité HDAC[144]. Une fois de plus, Kanthasamy *et al*. ont donc montré qu'une hyperacétylation des histones était impliquée dans les mécanismes de neurotoxicité des pesticides. Toujours en rapport avec la neurodégénérescence, une étude récente publiée par Maloney *et al*. montre que certaines perturbations du code histone induites par les pesticides peuvent augmenter la production de la protéine ß-amyloïde, l'une des deux marques histopathologiques de la maladie d'Alzheimer[145]. Ces perturbations épigénétiques

induites par l'environnement qui peuvent paraître temporaires deviennent en réalité latentes conduisant ainsi au développement de maladies neurodégénératives[146].

III.2.6 Les hydrocarbures aromatiques polycycliques

Les hydrocarbures aromatiques polycycliques (HAP) constituent une classe de composés organiques largement présents dans l'environnement. Ils résultent de la combustion incomplète de matière organique. Parmi les sources de contamination majeures chez l'Homme, on retrouve la fumée de cigarette mais également l'alimentation (graines, légumes, viandes grillées)[147]. Ces composés sont connus pour leurs propriétés carcinogènes dues à la formation par les CYP de métabolites électrophiles très réactifs. Le benzo[*a*]pyrène (B[*a*]P) est le chez de file de ces HAP. C'est un des toxiques majeurs présents dans la fumée de cigarette. Ses effets toxiques connus sont médiés par son métabolite réactif, le benzo[a]pyrène-*trans*-7,8-dihydrol-9,10-époxide (BPDE). Chez une lignée cellulaire de cancer du sein, un changement global du niveau d'acétylation de H3K9 a été observé au niveau des régions promotrices des gènes après une exposition de plusieurs jours au B[*a*]P[148]. Ce changement global du niveau d'acétylation s'est révélé être corrélé au niveau de transcription de certains gènes chez cette même lignée cellulaire. Même si une perturbation de l'épigénome par les HAP semble évidente sachant qu'ils induisent une modulation de l'expression de certains gènes, peu d'études sont parvenues à corréler ces changements à l'échelle des histones.

De manière plus générale, l'exposition à certains toxiques a clairement été reliée à une perturbation de l'épigénome et plus particulièrement du code histone (figure 33). Le code histone est aujourd'hui clairement établi comme étant un acteur nécessaire à la mise en place et au maintien du programme génétique normal. Sa susceptibilité vis-à-vis de l'environnement n'est plus à démontrer, ainsi que son implication dans l'étiopathogénie de nombreuses maladies. Il devient donc de plus en plus important de déchiffrer le code histone pour expliquer, voire prédire, le développement de certaines pathologies chroniques, et également pour évaluer les risques inhérents à une exposition à certains toxiques.

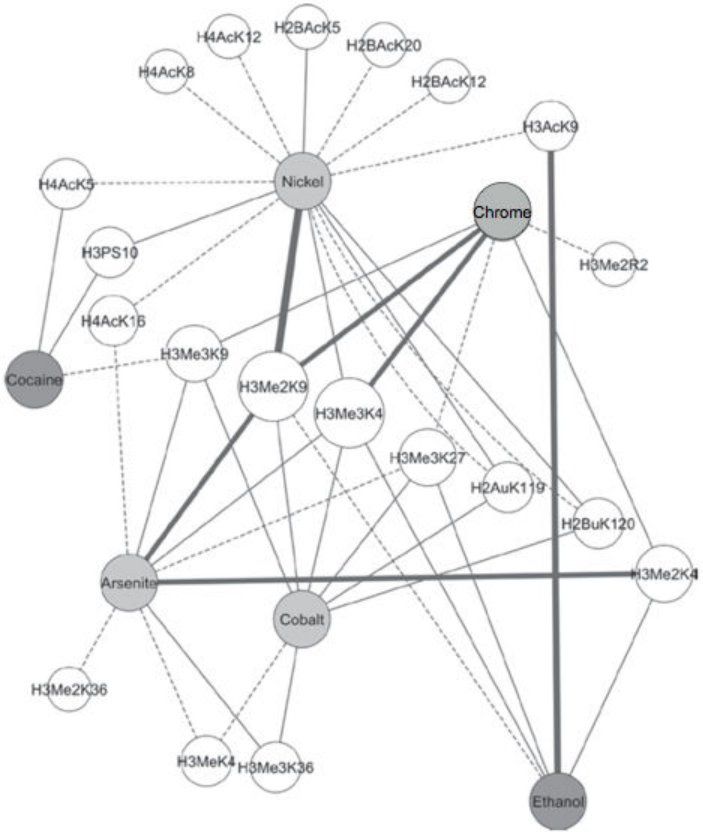


Figure 33 : représentation des interactions entre une exposition environnementale à certains toxiques et les modifications post-traductionnelles des histones. Les cercles gris représentent les toxiques considérés. Les cercles blancs représentent les modifications spécifiques d'histones. Les lignes pleines indiquent que la modification est augmentée par le toxique, et les lignes pointillées indiquent que la modification est diminuée par le toxique. Plus le trait est épais plus le nombre de publications confirmant la relation établie entre un toxique et une modification est élevé. D'après[149].

Face à la complexité du code histone, il convient de recourir à des méthodes analytiques suffisamment sensibles, spécifiques et résolutives pour tenter de le déchiffrer et d'obtenir une information pertinente. Les enjeux de la caractérisation du code histone sont donc multiples : comprendre, diagnostiquer, et prédire toutes les pathologies pouvant impliquer une perturbation épigénétique, mais également développer de nouvelles stratégies thérapeutiques ciblées visant à moduler l'épigénome. Au fil du temps, la spectrométrie de masse s'est imposée comme un des outils les mieux adaptés pour relever ce défi et déchiffrer le code histone.

IV. Méthodes analytiques pour déchiffrer le code histone

Face à l'enjeu que représentent les histones et leurs modifications post-traductionnelles pour la santé humaine, il est indispensable de disposer d'outils analytiques robustes permettant l'identification et la quantification de l'ensemble de ces marques épigénétiques. Il existe actuellement deux grands types de méthodes exploratoires utilisées pour l'étude du code histone : les méthodes immunologiques basées sur l'utilisation d'anticorps et les méthodes basées sur l'utilisation de la spectrométrie de masse. Nous verrons au cours de chapitre quels sont les avantages et les inconvénients de chacune de ces approches. Ce rapide passage en revue des méthodes disponibles pour explorer le code histone nous amènera à comprendre pourquoi, dans un contexte de toxicologie, nous avons cherché à développer une stratégie alternative capable d'identifier une perturbation du code histone à l'échelle globale.

IV.1 Méthodes immunochimiques

Traditionnellement, la recherche de modifications post-traductionnelles d'histones se fait à l'aide de méthodes immunologiques basées sur la reconnaissance spécifique d'un résidu modifié par un anticorps[150]. Le Western Blot ou immunotransfert à partir d'un gel d'électrophorèse est une des techniques de base en biochimie permettant de détecter la présence de certains résidus modifiés et d'obtenir une information semi-quantitative sur leur abondance. Les protéines éventuellement dénaturées sont séparées en fonction de leur masse sur un gel de polyacrylamide à partir duquel elles seront ensuite transférées sur une membrane (en nitrocellulose ou en polyfluorure de vinylidène) puis immunorévélées par exposition à un anticorps primaire spécifique de la modification ou du variant d'histone recherché. La révélation se fait généralement à l'aide d'un anticorps secondaire dirigé contre un épitope espèce-spécifique de l'anticorps primaire qui, étant lié à une enzyme, émettra un signal colorimétrique ou photométrique dont l'intensité sera proportionnelle à la quantité de protéine. Ces réactions anticorps-antigènes peuvent également être utilisées dans d'autres méthodes plus élaborées comme l'immunoprécipitation de chromatine (*Chromatin ImmunoPrecipitation*, ChIP). Cette technique consiste à utiliser un anticorps spécifique d'un type

d'histone ou d'un résidu modifié afin de faire précipiter la protéine cible liée à l'ADN préalablement digéré qui sera également identifié par séquençage. Elle présente l'énorme avantage de pouvoir associer un type d'histone modifiée avec une région spécifique du génome.

Bien que largement utilisées, ces méthodes immunologiques présentent plusieurs biais qui rendent discutables les informations qu'elles fournissent. Tout d'abord, il existe un véritable problème de réaction croisée des anticorps du à la grande similarité structurale des différents antigènes. Par exemple, un anticorps spécifique dirigé contre un site modifié pourra reconnaitre d'autres sites portant la même modification. Le deuxième inconvénient concerne le masquage possible d'un épitope par une modification post-traductionnelle voisine qui empêcherait sa reconnaissance par l'anticorps dirigée contre elle. Ce phénomène est particulièrement fréquent dans le cas des histones qui portent de très nombreuses modifications post-traductionnelles[151]. Enfin, le recours à un anticorps spécifique pour chaque résidu modifié oblige à avoir un certain *a priori* sur la nature des modifications que l'on recherche, et à les étudier une à une de façon indépendante, perdant ainsi totalement l'information combinatoire. Ceci représente un inconvénient majeur lorsque l'on cherche éventuellement de nouveaux sites de modifications ou que l'on cherche à explorer à l'aveugle l'ensemble des histones et de leurs modifications. De plus, nous avons vu que ces modifications n'étaient pas indépendantes les unes des autres et qu'elles devaient être considérées non pas individuellement mais comme un tout. Ces méthodes immunochimiques restent donc relativement peu informatives et trop ciblées lorsque l'on cherchera à explorer le code histone dans sa globalité.

Comparativement, les méthodes basées sur la spectrométrie de masse dont nous allons parler fournissent une information plus globale et plus détaillée que les méthodes immunologiques. Elles offrent la capacité unique de comparer qualitativement et semi-quantitativement les contenus en protéines provenant de différentes conditions, comme lorsque l'on compare des cellules saines et des cellules exposées à un toxique. Enfin, ces techniques à haut débit permettent dans certains cas d'appréhender l'aspect combinatoire du code histone.

IV.2 Stratégies en analyse protéomique

IV.2.1 Définition de la protéomique

Le terme protéome a été proposé pour la première fois par Mark Wilkins en 1994. Il définissait alors l'ensemble des protéines codées par le génome[152]. La protéomique est la discipline qui s'intéresse à l'étude du protéome, englobant ainsi l'identification, la caractérisation et la quantification de l'ensemble des protéines exprimées dans une cellule donnée, à un temps *t* et dans des conditions environnementale bien précises. La protéomique concerne tant les protéines et leurs isoformes que les modifications post-traductionnelles en passant par les interactions protéine-protéine. Tous les aspects du code histone sont donc appréhendés par la protéomique. De par la complexité, la diversité et la dynamique du protéome, l'analyse protéomique impose de recourir à un outil analytique à la fois résolutif, sensible et spécifique : la spectrométrie de masse [153]. De plus, l'analyse protéomique générant un volume important de données complexes elle est aujourd'hui indissociable de l'utilisation d'outils bioinformatiques adaptés.

IV.2.2 La spectrométrie de masse pour l'analyse des protéines

IV.2.2.1 Principe de fonctionnement d'un spectromètre de masse

La spectrométrie de masse est une méthode d'analyse apparue avec les découvertes de J.J. Thomson et F.W. Aston au début du XX^{e} siècle. Le premier instrument n'a été commercialisé qu'en 1942 aux Etats-Unis, et il faudra attendre la fin des années 1980 pour que l'instrumentation permette l'analyse des macromolécules, dont les protéines. La spectrométrie de masse est fondée sur la mesure en phase gazeuse du rapport masse-sur-nombre de charges noté *m*/*z* de molécules ionisées présentes dans un échantillon. Classiquement, un spectromètre de masse est composé de trois éléments : une source d'ionisation, un analyseur de masse et un détecteur (figure 34). La source d'ionisation permet à la fois l'ionisation et le passage à l'état gazeux des molécules. L'analyseur sépare sous vide les molécules ionisées en fonction de leurs valeurs *m*/*z*. Enfin le détecteur

reçoit un flux d'ions provenant de l'analyseur et émet un signal proportionnel au courant ionique.

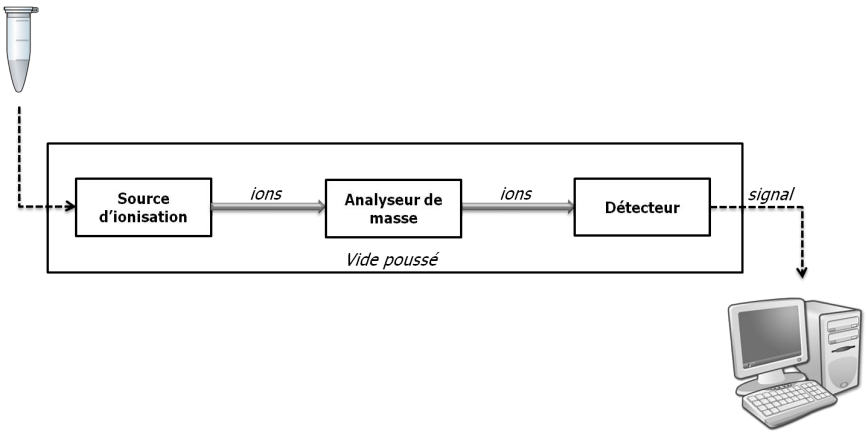


Figure 34 : schéma du principe de fonctionnement d'un spectromètre de masse.

IV.2.2.2 Les sources d'ionisation en protéomique

Les modes d'ionisation pour l'étude des protéines doivent tenir compte de leur caractère thermolabile et permettre de les volatiliser dans la source sans les fragmenter ni les dégrader. Deux techniques d'ionisation douce sont ainsi apparues dans les années 1980 et ont rendu possible l'analyse de protéines intactes en phase gazeuse : la désorption laser assistée par matrice (MALDI) et l'électronébulisation (ESI).

IV.2.2.2.1 La désorption laser assistée par matrice (MALDI)

La désorption laser assistée par matrice appelée communément MALDI (*Matrix-Assisted Laser Desorption-Ionization*) est un mode d'ionisation introduit en 1988 par Karas et Hillenkamp[154]. Cette technique fait appel à une matrice solide aromatique (M) introduite en large excès dans laquelle est dispersé l'échantillon à analyser (A). L'échantillon et la matrice sont co-cristallisés sur une plaque porte-échantillon généralement en inox qui est ensuite irradiée sous vide (≈ 10^{-7} mbar) par un faisceau laser pulsé de longueur d'onde donnée, majoritairement dans l'UV (figure 35). La matrice absorbe à la longueur d'onde du laser et entraîne l'analyte dans un plasma d'expansion ou il sera ionisé après excitation électronique et thermique selon la réaction de photoionisation :

$$M + h\nu \rightarrow M^{\bullet +} + e^{-}$$

Il s'ensuit des réactions de transfert de radical $H^{\bullet}$ et de protons H^{+} entre des radicaux matrices, des molécules de matrice et des molécules d'analyte A :

$$M^{\bullet+} + M \rightarrow [M+H]^{+} + [M-H]^{\bullet}$$
$$[M+H]^{+} + A \rightarrow M + [A+H]^{+}$$

Les ions d'analyte ainsi obtenus sont accélérés par une différence de potentiel jusqu'à l'analyseur.

Lors d'une analyse par spectrométrie de masse MALDI réalisée en mode positif, ce sont majoritairement des molécules monochargées protonées $[A+H]^{+}$ qui sont formées. D'un point de vue purement pratique, l'efficacité d'ionisation dépendra de plusieurs paramètres[155,156] : le choix de la matrice (acide α-cyano-4-hydroxycinnamique, acide 2,5-dihydroxybenzoique ou acide sinapinique par exemple), le mode de dépôt pour la co-cristallisation ou encore le pH des solutions. L'ionisation MALDI présente l'avantage d'être relativement tolérante aux sels.

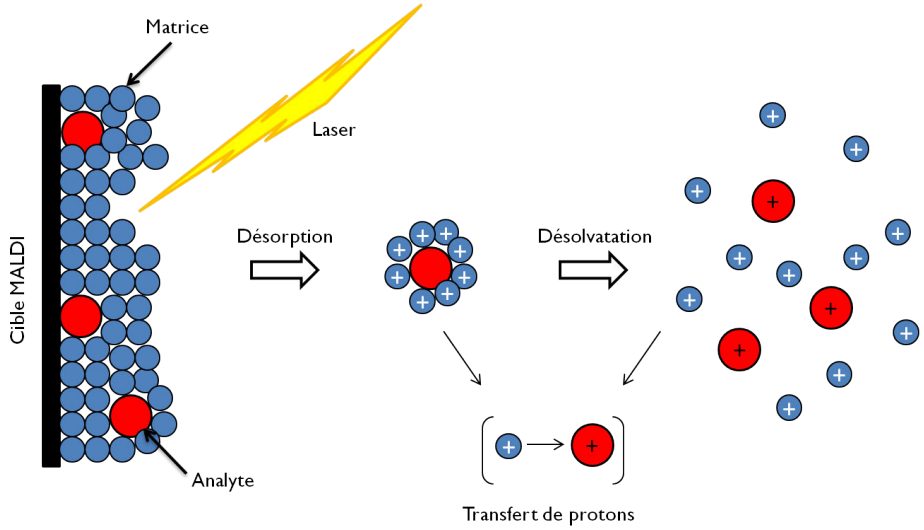


Figure 35 : principe de l'ionisation MALDI. Le processus se déroule en trois étapes : les molécules de matrice sont excitées par les photons du laser, puis les molécules de matrice et d'analytes sont désorbées et ionisées en phase gazeuse. Adapté de[157].

IV.2.2.2.2 L'électronébulisation (ESI)

L'électronébulisation ou ionisation *electrospray* est notée ESI (*electrospray ionization*). Cette méthode d'ionisation douce a été proposée pour la première fois par Dole et ses collaborateurs[158] dans les années 1970. Elle n'a cependant été

appliquée aux protéines qu'à partir de 1985 à la suite des travaux de Fenn sur les états de charge multiples des grosses molécules[159]. L'électronébulisation consiste à infuser à pression atmosphérique un échantillon en solution dans un capillaire soumis à un fort champ électrique. L'infusion se fait le plus souvent à un débit de l'ordre de la dizaine de microlitres par minute ($\mu L.min^{-1}$). Le champ électrique intense (10^6 V/m) créée par l'application d'une différence de potentiel entraîne la polarisation de la phase liquide et la séparation des charges positives et négatives. Pour l'analyse des protéines, le capillaire métallique est chargé le plus souvent positivement et joue le rôle de l'anode, neutralisant ainsi les charges négatives. Les charges positives se retrouveront donc en excès à l'extrémité du capillaire entraînant une distorsion du liquide que l'on appelle cône de Taylor[160]. Ce cône va s'étirer jusqu'à se dissocier en minces gouttelettes qui rencontreront en phase gazeuse un flux d'azote chaud à contre-courant conduisant à l'évaporation progressive des molécules de solvant (figure 36). Cette évaporation progressive du solvant provoque une diminution de la taille des gouttelettes et une augmentation parallèle de densité de charge[161]. Dès lors que le rayon de la gouttelette devient inférieur au rayon critique de Rayleigh, elle devient instable et subit une explosion coulombienne générant des gouttelettes filles de deuxième génération qui subiront à leur tour le même processus jusqu'à obtenir des ions multichargés intégralement désolvatés[162].

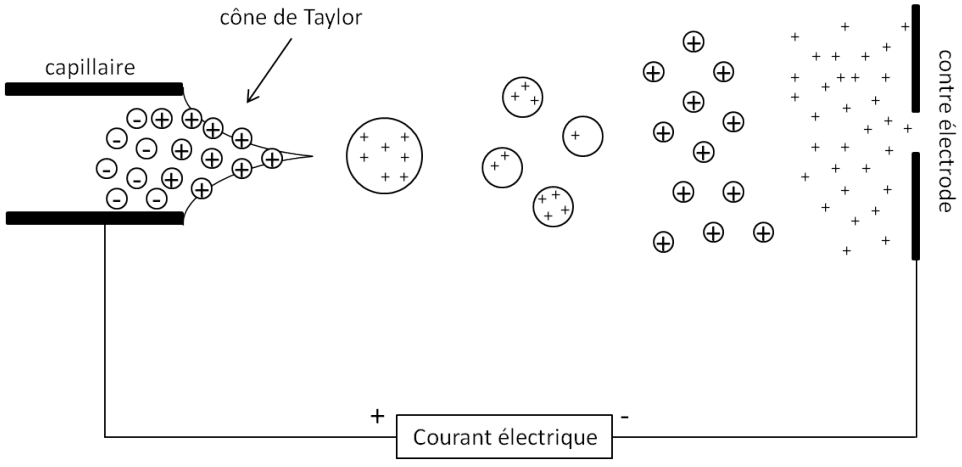


Figure 36 : représentation schématique de la production d'ions par électronébulisation. Adapté de[157].

Le processus d'électronébulisation permet au détecteur du spectromètre de masse de mesurer un courant ionique qui est fonction non pas du débit d'infusion

mais de la concentration des analytes présents dans l'échantillon. Lorsque l'on souhaite réaliser des analyses à partir de faibles quantités d'échantillons, il peut être intéressant pour gagner en sensibilité de travailler à des débits plus faibles et d'utiliser des solutions plus concentrées. Des sources *micro-electrospray*[163] et *nano-electrospray*[164] ont ainsi été développées qui permettent d'atteindre des débits respectifs de l'ordre de quelques $\mu L.min^{-1}$ à quelques $nL.min^{-1}$.

En spectrométrie de masse ESI, les grosses molécules possédant plusieurs sites ionisables telles que les protéines, et les histones en particulier, produisent des ions multichargés. Ceci présente comme avantage de permettre l'analyse de molécules de masse moléculaire importante par des analyseurs dont la gamme de masse nominale est faible. Les spectres de masse ESI des protéines représentent donc une distribution statistique de pics correspondant aux ions moléculaires multichargés $[M+zH^+]^{z+}$ (figure 37).

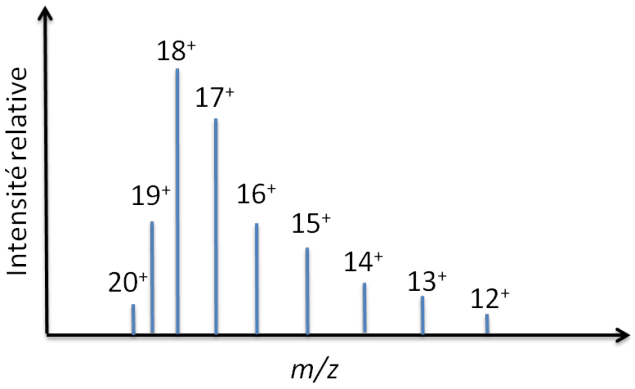


Figure 37 : représentation schématique d'un spectre ESI d'une protéine entière. Les chiffres indiqués représentent le nombre de charge z de chacun des ions moléculaires.

Chaque pic apparent sur un spectre ESI correspondant à un rapport *m/z*, il sera possible d'en déduire la masse *m* de la molécule en déterminant le nombre de charge z. Pour cela, il suffira de considérer les rapports *m/z* de deux pics adjacents consécutifs X et Y selon la formule : $z_Y = \frac{(X-1)(Y-1)}{X-Y}$ où X correspond au rapport *m/z* d'un pic noté X et Y au rapport *m/z* du pic suivant X. Il existe actuellement dans les logiciels commerciaux des algorithmes de déconvolution permettant de déterminer automatiquement la masse moléculaire d'une protéine en transformant les pics multichargés en un pic monochargé à partir de son spectre ESI. Lorsque

l'on travaille à haute résolution, un niveau supplémentaire d'information vient s'ajouter à la distribution des pics multichargés. Chaque ion multichargé est scindé en plusieurs pics correspondant à sa distribution isotopique. Dans ce cas, la distance observée entre deux pics adjacents d'un même massif isotopique correspond à $\frac{1}{z}$ et permet de déterminer directement l'état de charge de l'ion.

IV.2.2.3 Les analyseurs

IV.2.2.3.1 Généralités

L'analyseur de masse est le principal responsable de la qualité des spectres obtenus lors d'une analyse. Il en existe plusieurs types qui fonctionnent selon des principes bien différents. Les analyseurs sont classiquement répartis en deux catégories : les analyseurs à faisceau d'ions qui offrent une analyse dans l'espace, et les analyseurs à piégeage d'ions qui offrent une analyse dans le temps. Un analyseur doit être choisi en fonction de l'information que l'on souhaite obtenir et de ses performances. Pour définir les performances d'un analyseur, plusieurs paramètres sont évalués : la gamme de masse, l'exactitude sur la mesure de masse, la résolution et la sensibilité. La résolution R d'un analyseur est sans doute le paramètre le plus important lorsque l'on s'intéresse à des mélanges complexes contenant des protéines très similaires, comme c'est le cas avec les histones. Elle représente la capacité d'un analyseur à séparer deux pics voisins de valeurs *m/z* très proches. Généralement, on considère que deux pics sont résolus si l'intensité de la vallée entre ces deux pics égale 10% de l'intensité du pic le plus faible (figure 38). Ainsi, si *Δm* est la plus petite différence de masse observée à 10% de vallée pour deux pics résolus de masse *m* et *m* + *Δm*, la résolution R est définie par le rapport : $R = \frac{m}{\Delta m}$. Il est également possible de calculer la résolution pour un pic isolé en prenant pour *Δm* la largeur du pic à 50% de son maximum. On parlera alors de résolution FWHM (*Full Width at Half Maximum*). Il sera fait référence uniquement à la résolution FWHM dans ce manuscrit. En pratique, si la résolution en masse d'un analyseur dépasse 5 000 FWHM celui-ci est dit à haute résolution.

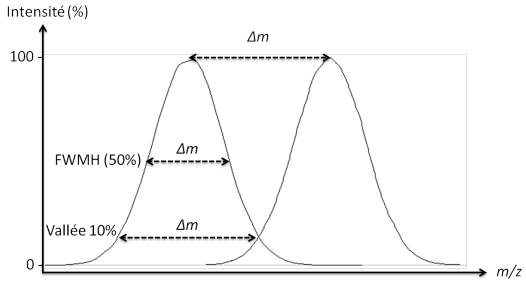


Figure 38 : schéma illustrant la notion de résolution et les deux façons usuelles de la calculer.

Il est très important de bénéficier d'une bonne résolution lorsque l'on travaille sur des protéines pour pouvoir déterminer l'état de charge des ions multichargés en ESI afin d'en déduire la masse moyenne, mais également si l'on souhaite avoir accès à la distribution isotopique.

Cependant la résolution d'un analyseur permet d'obtenir des mesures de masse précises mais pas obligatoirement exactes. Le deuxième paramètre à prendre en compte est donc l'exactitude sur la mesure de masse, qui dépend de la qualité de l'étalonnage. Il s'agit de la capacité de l'analyseur à mesurer un rapport *m/z* le plus proche possible de la valeur théorique. Elle est définie par le rapport $\frac{\Delta m_e}{m} \times 10^6$ où Δm_e représente l'erreur relative commise sur la mesure de masse théorique *m* et s'exprime en partie par millions (ppm). Enfin, la sensibilité d'un analyseur correspond à la quantité minimale de molécules d'analyte qu'il est capable de détecter, et la gamme de masse correspond au rapport *m/z* maximal détectable.

Au cours de ce travail de thèse, différents instruments ont été utilisés et seuls les principes des analyseurs constituant ces appareils seront décris (quadripôle, analyseur à temps de vol, instruments hybrides). À titre d'exemple, le tableau 6 résume les performances des analyseurs qui seront étudiés au cours de ce chapitre.

Tableau 6 : comparaison des performances théoriques des différents analyseurs de masse composant les instruments utilisés au cours de ces travaux de thèse.

Analyseur	Résolution (FWHM) à la masse *m*/*z* 400	Limite en masse (*m*/*z*)	Exactitude en masse (ppm)
Quadripôle (Q)	3 000	4 000	200
Temps de vol (TOF)	8 000 – 60 000	> 1 000 000	2-10 (après étalonnage interne)

IV.2.2.3.2 Le quadripôle

L'analyseur quadripolaire ou quadripôle (Q) est un analyseur à basse résolution qui a été décrit pour la première fois en 1953 par Paul et Steinwedel[165]. Il est formé de quatre barres métalliques parallèles de section interne idéalement hyperbolique. Le potentiel électrique $+\Phi_0$ ou $-\Phi_0$ des barres opposées est identique tandis que celui des barres adjacentes est opposé (figure 39). Ce potentiel est composé d'une tension continue U et d'une tension alternative V (radio-fréquence) qui définissent la relation $\Phi_0 = \pm(U - V \times \cos \omega t)$ où ω représente la fréquence des signaux alternatifs. Les ions formés dans la source entrent dans l'analyseur et subissent l'effet du champ électrostatique bidimensionnel (x et y). Ils adoptent alors une trajectoire plus ou moins stable suivant l'axe z qui dépend des valeurs de U et de V et qui est calculée à l'aide des équations de Mathieu. Les ions qui adoptent une trajectoire instable viennent terminer leur course et se décharger sur les barres du quadripôle. En balayant les tensions, il est possible de faire traverser l'analyseur à chaque ion successivement et d'enregistrer un spectre de masse. Le quadripôle est donc un analyseur à balayage qui se base sur la stabilité de la trajectoire des ions pour les séparer selon leur rapport *m*/*z*.

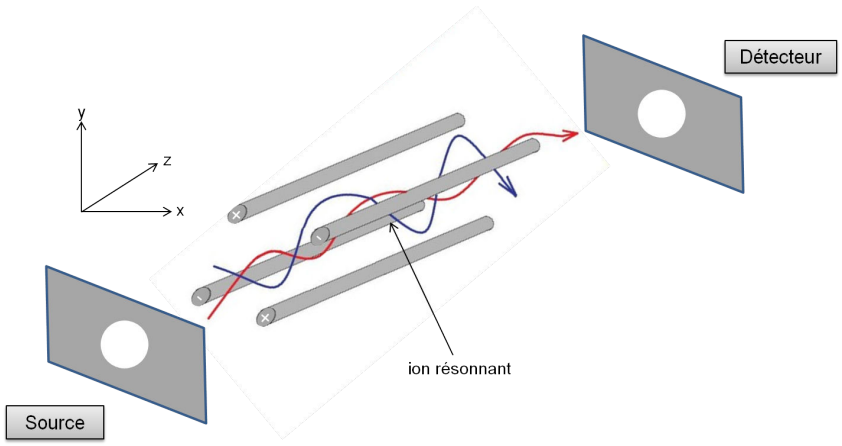


Figure 39 : schéma d'un analyseur quadripolaire illustrant la trajectoire des ions selon l'axe z.

IV.2.2.3.3 L'analyseur à temps de vol

L'analyseur à temps de vol (*Time-of-Flight*, TOF) repose sur un principe de séparation des ions découvert dans les années 1950[166]. Cependant, il a fallu attendre l'apparition de la source MALDI pour que cet analyseur soit largement utilisé[154]. Son principe de fonctionnement est relativement simple et consiste à mesurer le temps mis par un ion pour parcourir une distance déterminée dans un tube de vol sous vide et libre de tout champ. A l'entrée du tube de vol, les ions sont accélérés par une différence de potentiel notée V et acquièrent une énergie cinétique E_c proportionnelle à leur masse selon la relation : $E_c = \frac{1}{2} mv^2 = zeV$ (m : masse de l'ion, z : nombre de charge, e : masse élémentaire). La mesure du temps d'arrivée t des ions au détecteur permet de calculer la valeur du rapport m/z selon l'équation : $t = \sqrt{\frac{m}{2zeV}} \times d$ (d : longueur du tube de vol). Les ions accélérés volent ainsi d'autant plus vite qu'ils sont plus légers.

En mode linéaire (figure 40a), la résolution en masse d'un analyseur TOF est seulement de quelques centaines pour un tube de vol de un à deux mètres de long. Ce manque de résolution s'explique par des phénomènes de variations de distribution temporelle, spatiale et cinétique des ions. Afin d'améliorer sensiblement cette résolution, un premier système de miroirs électrostatiques appelé réflectron (figure 40b) a été mis au point. Il permet de corriger la dispersion cinétique en refocalisant les ions de même rapport m/z et également d'allonger le temps de vol. Un deuxième système d'extraction retardée des ions a

été mis au point pour corriger les dispersions spatiales et temporelles survenant en source. Ce système permet de synchroniser le départ des ions vers le tube de vol[167]. La combinaison de ces deux systèmes permet d'atteindre pour des rapports *m/z* inférieurs à 10 000 des résolutions de l'ordre de 20 000 sur les instruments commerciaux.

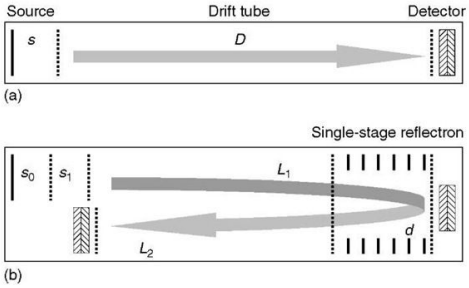


Figure 40 : schéma d'un analyseur TOF en mode linéaire (a) et en mode réflectron (b). D'après[167].

Classiquement, les analyseurs TOF sont couplés à des sources pulsées telles que le MALDI. Cependant, l'apparition des analyseurs TOF à injection orthogonale (oaTOF) a permis leur couplage avec des sources continues telles que l'ESI. Dans ce cas de figure, le faisceau d'ions incidents est perpendiculaire au tube de vol et est dévié vers celui-ci par un champ électrostatique appliqué au niveau d'un accélérateur orthogonal (figure 41). Cette accélération orthogonale imposée par le potentiel pulsé réduit la dispersion cinétique des paquets d'ions et permet de contrôler indépendamment la production des ions de la source à l'accélérateur orthogonal et leur analyse[168].

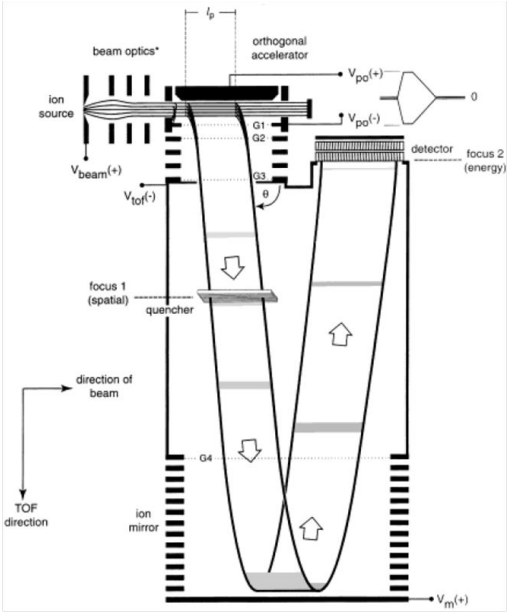


Figure 41 : schéma d'un analyseur TOF à injection orthogonale. D'après[168].

Les analyseurs TOF à injection orthogonale permettent d'atteindre des hautes fréquences d'acquisition en adéquation avec la courte durée de vol des ions dans le tube de vol (de l'ordre de 50 µs). De plus, ils peuvent atteindre jusqu'à 60 000 de résolution.

IV.2.2.3.4 Analyseurs en tandem : le spectromètre de masse hybride Q-TOF

L'utilisation de deux analyseurs en tandem au sein d'un même spectromètre de masse permet d'augmenter la sélectivité et la sensibilité de la détection. Il existe plusieurs types d'analyseurs en tandem, mais il ne sera question ici que des analyseurs en tandem de type Q-TOF. Ces analyseurs en tandem permettent de réaliser des expériences de fragmentation des ions dans une cellule de collision appelées analyses MS/MS dont nous parlerons lors du prochain chapitre. Le rapport *m*/*z* des ions précurseurs est mesuré par le quadripôle (MS1) tandis que les rapports *m*/*z* des ions fragments sont mesurés par l'analyseur TOF à injection orthogonale (MS2). Au cours de ces travaux, le spectromètre de masse SYNAPT G2 HDMS de la société Waters a été majoritairement utilisé. Il s'agit d'un instrument hybride Q-TOF à haute résolution composé d'une source ESI, d'un quadripôle, de deux cellules de collision, d'une cellule de mobilité ionique, d'un analyseur TOF à injection orthogonale et d'un détecteur (figure 42).

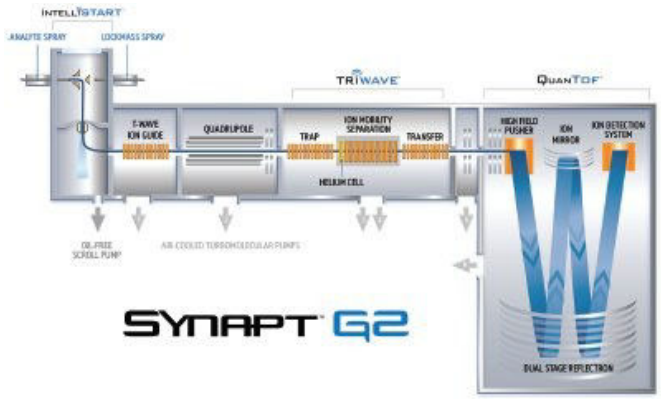


Figure 42 : schéma du spectromètre de masse SYNAPT G2 HDMS (Waters Corporation, Manchester, UK).

Le faisceau d'ions est focalisé à l'aide d'un guide d'ions à ondes progressives (*T-wave ion guide,* TWIG) avant leur entrée dans le quadripôle. Contrairement aux quadripôles présents dans certains appareils tels que les triple quadripôles (TQ), celui intégré au SYNAPT G2 n'effectue pas de mesure en vue d'acquérir un spectre MS1 mais sert soit de guide d'ions en mode MS soit de filtre en mode MS/MS. A la sortie de ce quadripôle, les ions entrent dans la partie du spectromètre appelée TriWave destinée à la fragmentation et à la mobilité ionique. Cette partie est constituée d'une première cellule TWIG trappe, d'une cellule de mobilité ionique à proprement parler et d'une seconde cellule TWIG de transfert. Les ions sont ensuite dirigés vers l'analyseur TOF à injection orthogonale possédant un réflectron à deux étages qui les renvoient vers le détecteur.

L'analyseur TOF présente la particularité de pouvoir être utilisé selon trois modes différents qui dépendent du rapport sensibilité/résolution recherché. En « mode sensibilité », la trajectoire des ions dans le tube de vol décrit un V, plafonnant ainsi la résolution à 10 000. En « mode résolution », les ions suivent toujours une trajectoire en V mais la résolution peut-être portée à 20 000. En « mode haute-résolution », les ions suivent une trajectoire en W dans le tube de vol, ce qui permet d'atteindre une résolution de 30 000. L'exactitude est également excellente, avec une erreur sur la mesure de masse généralement inférieure à 10 ppm. En effet, la stabilité de l'étalonnage de l'analyseur TOF est garantie par l'infusion constante d'un étalon de masse connue (*Lockspray*®). Le SYNAPT G2 est donc un instrument performant doté d'une certaine flexibilité de

ses paramètres, ce qui permet de réaliser une large gamme d'expériences sur des molécules de nature très différente (protéines, lipides, petites molécules chimiques).

IV.2.3 Les techniques séparatives

De manière générale, la complexité des mélanges de protéines issues de milieux biologiques rend nécessaire le recours à une séparation préalable des protéines avant leur analyse par spectrométrie de masse. Ceci est d'autant plus vrai dans le cas des histones du fait de la diversité des variants rencontrés. La séparation des différentes formes d'histones représente un réel défi tant elles ont des masses moléculaires et des points isoélectriques proches. Les techniques couramment utilisées pour séparer les histones peuvent être réparties en deux classes : les techniques basées sur l'utilisation de gels et celles basées sur l'utilisation de colonnes chromatographiques. Le choix de la technique séparative dépend du type d'information que l'on souhaite obtenir et de la complexité des mélanges. Ainsi, des techniques différentes seront employées selon que l'objectif est d'analyser en ligne les histones ou de collecter des fractions purifiées. Dans le cas de l'analyse des histones, ce choix reste également limité par la compatibilité de la technique avec l'analyse par spectrométrie de masse en aval. Dans ce chapitre, nous nous limiterons aux techniques fréquemment utilisées pour séparer les histones avant leur analyse par spectrométrie de masse.

IV.2.3.1 L'électrophorèse sur gel de polyacrylamide en conditions dénaturantes

Le gel d'électrophorèse en condition dénaturante SDS-PAGE (*sodium dodecylsulfate-polyacrylamide gel electrophoresis*) est une technique qui permet de séparer des protéines en fonction de leurs masses moléculaires (figure 43).

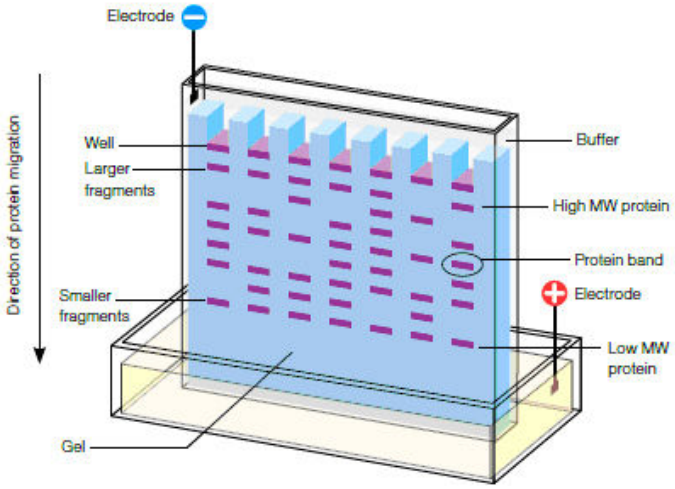


Figure 43 : représentation schématique de la séparation de protéines par électrophorèse sur gel de polyacrylamide. Selon le système discontinu de Laemmli[169], le gel se compose de deux parties : une partie supérieure étroite appelée gel de concentration (« *stacking* ») à larges pores contenant du tampon Tris-HCl à pH 6,8 et une partie majoritaire appelée gel de séparation (« *resolving* ») contenant du tampon Tris-HCl à pH 8,8. D'après http://www.bio-rad.com/fr-fr/applications-technologies/protein-electrophoresis-methods.

Cette technique, relativement simple, peut être utilisée pour purifier des échantillons complexes avant leur analyse par spectrométrie de masse. Le pourcentage en polyacrylamide du gel détermine sa résolution et est à choisir en fonction de la gamme de masses moléculaires des protéines d'intérêt (généralement entre 5 et 15%). Ainsi, plus la concentration en polyacrylamide est élevée, plus les pores du gel seront étroits. La présence de dodécylsulfate de sodium (SDS) entraîne une dénaturation des protéines qui permet de gommer les forces de freinage lors de la migration à travers les pores du gel (figure 44). Toutes les protéines migrent vers l'anode sous l'effet d'un champs électrique, et le ratio établi de 1,4 grammes de SDS par gramme de protéines leur confère une charge négative permettant de s'affranchir du facteur charge propre[169].

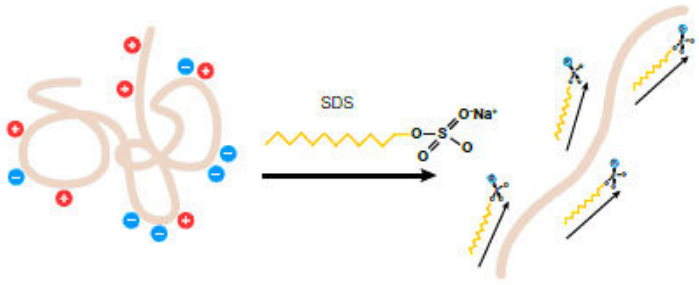


Figure 44 : effets du SDS sur la conformation et la charge d'une protéine. D'après Bio-Rad.

La distance de migration des protéines sera donc proportionnelle à leur masse moléculaire. L'utilisation de protéines de masses moléculaires connues en tant que marqueurs permet d'évaluer la masse moléculaire des protéines étudiées en comparant leur distance de migration sur le gel.

La sensibilité de cette technique dépend du type de détection utilisé pour révéler les protéines. Les deux principaux colorants utilisés sont le bleu de Coomassie et le nitrate d'argent. La coloration du gel au bleu de Coomassie colloïdal se fait par l'intermédiaire de réactions non covalentes (adsorption) entre le réactif et les groupements amine des acides aminés basiques ou aromatiques[170]. Cette coloration est relativement peu sensible (de l'ordre de 50 à 100 ng) mais linéaire. L'intensité de la coloration peut être artificiellement augmentée dans le cas de protéines très basiques telles que les histones, ce qui représente un biais lorsqu'on souhaite comparer leurs abondances relatives. La coloration au nitrate d'argent est, elle, beaucoup plus sensible[171] et repose sur la réduction de l'ion argent par les groupes sulfhydrile et carboxylique des chaines latérales des protéines. L'ion argent réduit ainsi formé est sensible à la lumière et permet la révélation des protéines présentes. Cependant, cette coloration n'est plus linéaire au-delà de certaines concentrations. Elle ne reflète donc pas la stœchiométrie des protéines révélées. De plus, il faut veiller à choisir le bon protocole lorsqu'on souhaite utiliser la spectrométrie de masse par la suite. Enfin, il existe d'autres colorations par fluorescence moins souvent utilisées mais tout aussi sensibles que le nitrate d'argent et qui restent compatibles avec l'utilisation de la spectrométrie de masse.

Dans le cas des histones, le pouvoir résolutif modeste de cette technique ne permet que de séparer les différents sous-types mais en aucun cas de séparer les variants d'un même sous-type d'histone. La présence de plusieurs protéines par bande oblige à coupler le SDS-PAGE avec une autre technique séparative bien plus performante : la chromatographie en phase liquide. Le SDS-PAGE ne représentera donc éventuellement qu'une étape de purification préliminaire permettant de séparer les différents sous-types d'histones avant par exemple une protéolyse enzymatique en gel.

IV.2.3.2 L'électrophorèse sur gel de polyacrylamide en présence d'acide acétique-urée

L'introduction des gels d'électrophorèse acide acétique-urée (AU-PAGE) et Triton acide acétique-urée (TAU-PAGE) a permis d'améliorer significativement la séparation des histones en gel. La présence d'acide acétique dans la composition des gels assure le maintien d'un pH acide (< 3) conduisant à la protonation des histones. Leur séparation peut ainsi se faire selon leur masse mais aussi selon leur charge [172]. Comme nous l'avons évoqué lors d'un précédent chapitre, l'acétylation et la phosphorylation des histones neutralisent directement ou indirectement une charge positive des protéines augmentant ainsi leur caractère acide. Le AU-PAGE permet donc de séparer des formes d'histones différentiellement acétylées ou phosphorylées. Cependant, malgré cette dimension de séparation supplémentaire, il peut arriver que différents variants acétylés de chacune des histones de cœur aient la même taille et la même charge, conduisant à un chevauchement de leur bande sur un gel AU-PAGE. Pour pallier cela, l'ajout d'un détergent non-ionique de type Triton X-100 a été proposé afin d'améliorer la séparation des différents variants et isoformes des histones de cœur. Les histones de cœur ont la faculté de se lier à ce détergent non-ionique qui, au-delà de permettre leur dénaturation, a pour conséquence d'augmenter leur taille, de réduire leur mobilité en gel et donc d'améliorer leur séparation. L'affinité des histones de cœur pour le Triton X-100 change considérablement en fonction des sous-types. Par ordre croissant d'affinité, on retrouve H4, H2B, H3 et H2A[173]. Cependant, il faut garder en tête que l'ajout de détergent non-ionique diminuera l'efficacité d'une éventuelle protéolyse en gel.

IV.2.3.3 La chromatographie en phase liquide

La chromatographie en phase liquide est devenue un outil incontournable lorsqu'il s'agit d'analyser des mélanges complexes de protéines ou de peptides protéolytiques. C'est une technique séparative qui possède de nombreux avantages et qui s'adapte parfaitement à l'étude des histones. Utilisée en amont de la spectrométrie de masse, elle permet notamment d'améliorer la détection des variants peu abondants et d'étendre la gamme dynamique. De manière très

basique, la chromatographie en phase liquide se compose d'une phase dite stationnaire qui est immobilisée dans une colonne et une phase dite mobile qui circule au contact de la phase stationnaire. L'échantillon à séparer est introduit en tête de colonne et la traverse de part en part en s'y adsorbant, entraîné par la phase mobile. Avec les sources d'ionisation modernes à pression atmosphérique, la spectrométrie de masse peut-être directement couplée en sortie de colonne chromatographique et utilisée comme détecteur. On parle ainsi de couplage LC-MS.

En protéomique, la technique la plus couramment utilisée est la chromatographie liquide haute performance (HPLC). Dans le cadre de ces travaux sur les histones, seule la chromatographie liquide à polarité de phase inversée (RP-HPLC) a été utilisée. Dans ce cas, la phase stationnaire est hydrophobe, à base de billes poreuses de silice dont le diamètre peut aller jusqu'à 5 µm et dont la taille des pores varie entre 60 et 300 Å. La surface de ces billes est greffée avec des chaines alkyles de longueur variable, généralement de 4 à 18 atomes de carbone (C_4 à C_{18}). La phase mobile est constituée d'un mélange d'eau avec un solvant organique polaire, classiquement l'acétonitrile, dans des proportions qui varient au cours du temps selon un gradient prédéfini. L'élution des histones se fera donc par ordre croissant d'hydrophobicité relative. Il existe également d'autres techniques analytiques ou préparatives telle que l'électrophorèse capillaire (CE) ou d'autres types de sélectivité telle que la chromatographie liquide d'interaction hydrophile (HILIC) qui ont été appliquées à la séparation des histones. Lindner *et al*. ont ainsi démontré l'apport significatif de la chromatographie HILIC pour la séparation des formes différentiellement acétylées des histones[174]. Bien que cette technique ait fait ses preuves, la chromatographie HILIC utilise des phases mobiles souvent riches en sels qui rendent son couplage en ligne avec la spectrométrie de masse difficile, et obligent à collecter les fractions en sortie de colonne avant dessalage. L'ajout de ces étapes supplémentaires représente un vrai inconvénient face au couplage en ligne RP-HPLC et spectrométrie de masse, qui reste aujourd'hui la méthode de choix pour l'étude des histones et de leurs modifications post-traductionnelles.

Il est possible d'améliorer la résolution et la sensibilité lors de la séparation chromatographique de protéines ou de peptides en jouant sur un certains nombre de paramètres autres que la composition ou le débit de phase mobile, parmi lesquels la longueur de la colonne, son diamètre interne, la granulométrie ou

encore la température. En spectrométrie de masse ESI, la mesure ne dépend pas du débit d'injection de l'échantillon mais de la concentration de ce dernier. Ainsi, des systèmes de chromatographie liquide ont-ils été développés qui permettent de réduire les débits de phase mobile et qui utilisent des colonnes de très faible diamètre interne. Ces techniques dites de nano-chromatographie (nano-HPLC) sont couplées à une source nanoESI et très largement appliquées à l'étude des peptides. Ces techniques permettent également de diminuer les volumes d'échantillons injectés, d'améliorer la séparation des peptides, de diminuer les effets de suppression d'ionisation et de réduire la consommation de phase mobile. Plus récemment, des systèmes chromatographiques permettant de travailler à très haute pression sont apparus. Il s'agit de la chromatographie ultra-haute performance UPLC (*Ultra Performance Liquid Chromatography*) développée par le constructeur Waters[175]. L'UPLC utilise des colonnes de granulométrie inférieure à 2 µm avec des vitesses linéaires de phase mobile plus importantes, ce qui permet d'améliorer significativement l'efficacité de la séparation et augmente le nombre de pics résolus par unité de temps (capacité de pics). La séparation chromatographique en UPLC se fait à très haute pression (jusqu'à 1000 bar) et améliore significativement le rapport signal sur bruit en diminuant la largeur des pics au profit d'une augmentation de leur hauteur. La durée des analyses est donc raccourcie par rapport aux méthodes chromatographiques classiques, sans sacrifier la résolution. L'intérêt de cette technique pour l'étude des histones a d'ailleurs été démontré par Contrepois *et al.* en 2010[176]. En parallèle, elle est de plus en plus utilisée dans d'autres disciplines telle que la métabolomique, où elle permet d'augmenter significativement le nombre de pics détectés[177] et de réduire la consommation de phases mobiles.

IV.2.4 Les stratégies d'analyse des histones et de leurs modifications post-traductionnelles par spectrométrie de masse

Nous venons de voir les différents outils analytiques utilisables pour l'étude des histones et de leurs modifications post-traductionnelles. Cependant, il existe plusieurs stratégies d'analyse combinant ces outils de différentes manières et faisant appel à la bioinformatique pour l'interprétation des données. De manière générale, toutes les stratégies employées ont recours à une séparation par

chromatographie liquide RP-HPLC utilisant des colonnes C_{18} ou plus rarement HILIC, couplée à la spectrométrie de masse ESI en tandem. La différence entre ces stratégies réside donc principalement dans la nature de l'échantillon injecté. Certaines utilisent les protéines intactes comme matériel de départ, d'autres des peptides de longueur variable issus de protéolyses enzymatiques. Comme nous le verrons au cours de ce chapitre, le type de matériel de départ conditionnera le choix du mode de fragmentation MS/MS afin d'obtenir le maximum d'information sur la nature et la localisation des modifications post-traductionnelles. Dans le cas des histones, le choix de la stratégie analytique se révèle compliqué de par la grande complexité du code histone et la forte identité de séquence entre les variants. Nous verrons que les stratégies classiques présentent parfois quelques inconvénients, ce qui nous a poussés à développer une stratégie alternative pour l'étude du code histone dans un contexte toxicologique.

IV.2.4.1 Identification par empreinte peptidique massique

L'amélioration des performances des instruments commerciaux ainsi que l'apparition des premières banques de séquences protéiques dans les années 1990 ont permis le développement de nouvelles stratégies pour l'identification des protéines. L'identification par empreinte peptidique massique (*Peptide Mass Fingerprint*, PMF) décrite pour la première fois en 1993[178,179], consiste à identifier les protéines présentes dans un échantillon par la mesure des rapports *m*/*z* des peptides générés par protéolyse enzymatique. Par l'intermédiaire de banques de séquences en ligne, les masses des peptides protéolytiques sont comparées à celles de peptides théoriques issus de la digestion *in silico* de toutes les protéines référencées pour un organisme donné (figure 45).

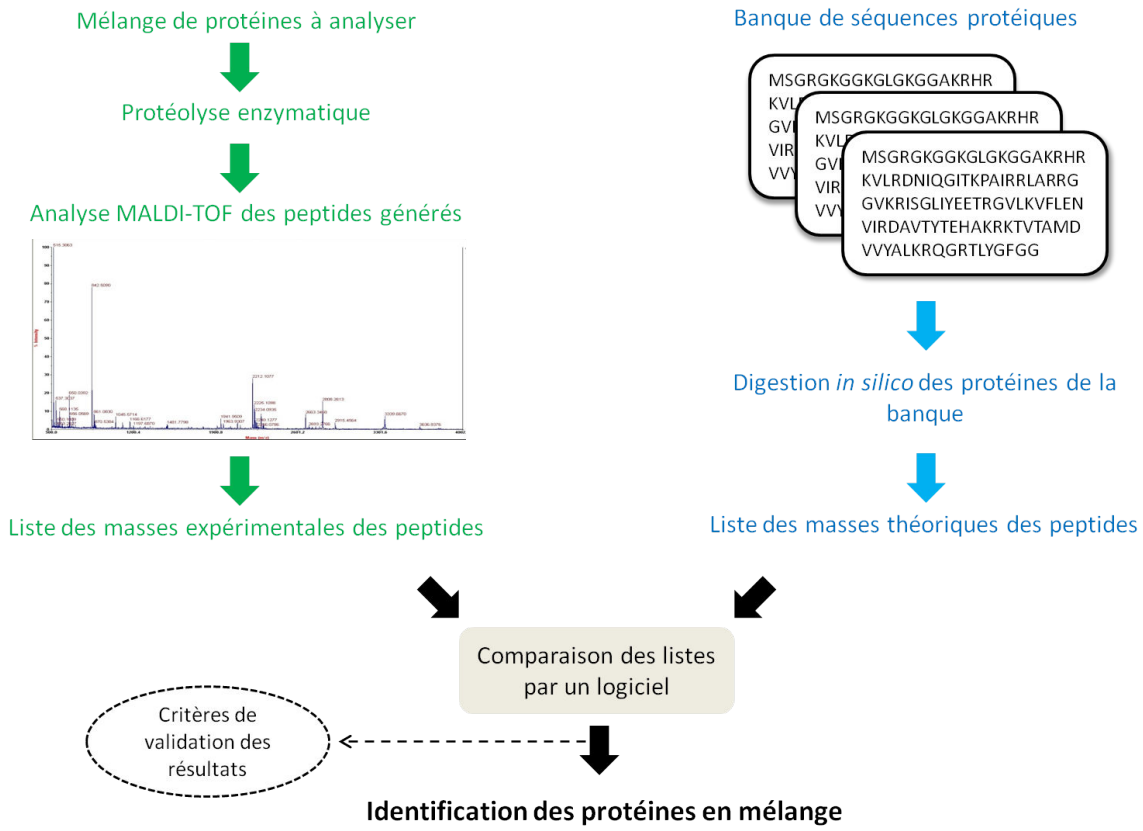


Figure 45 : stratégie d'identification de protéines par empreinte peptidique massique (PMF).

Grâce à sa bonne tolérance aux sels et aux détergents, sa simplicité d'utilisation et à ses bonnes performances, la spectrométrie de masse MALDI-TOF s'est imposée comme l'outil de choix pour l'approche PMF. Elle permet ainsi une identification rapide et assez précise du contenu en protéines d'échantillons relativement peu complexes. La validation de l'identification de protéines par PMF répond à quelques critères simples, à savoir le score de la protéine calculé à partir d'algorithmes propres à chaque logiciel, le nombre de peptides identifiés et appartenant à la protéine (couverture de séquence peptidique), et enfin les écarts entre les masses des peptides expérimentaux et théoriques. Lorsqu'elle est appliquée à l'étude du code histone, cette stratégie atteint rapidement ses limites. Bien que les instruments MALDI-TOF fournissent des informations précises sur la masse des peptides présents, ils ne possèdent pas la résolution suffisante pour distinguer des peptides isobares n'ayant pas la même séquence en acides aminés. Les instruments MALDI-TOF sont également soumis au phénomène de suppression spectrale qui correspond à l'écrasement du signal de certains peptides du à une efficacité d'ionisation séquence dépendante. La complexité du code histone ainsi

que son aspect combinatoire ne peuvent donc pas être appréhendés par ce type d'approche. De manière générale, lorsque le mélange à analyser est trop complexe, la superposition des empreintes peptidiques sur le même spectre rendra impossible une identification fiable. Face au nombre croissant d'entrées dans les banques de données, la masse seule d'un peptide ne suffit pas à garantir son identité. A côté de cela, un autre problème concerne la taille des peptides générés par protéolyse enzymatique. Dans la majorité des cas, l'enzyme de choix pour générer des peptides est la trypsine. Cette endoprotéase hydrolyse la liaison peptidique en C-ter d'une arginine ou d'une lysine. Nous avons évoqué lors d'un chapitre précédent la richesse des histones en résidus basiques arginine et lysine. Une trypsinolyse des histones générera donc de nombreux peptides de longueur inférieure à sept acides aminés. Ces nombreux petits peptides souvent non spécifiques augmenteront significativement la complexité du mélange et empêcheront de déterminer la nature des variants présents ainsi que leur stœchiométrie. La faible spécificité de l'approche PMF utilisant des instruments MALDI-TOF explique que son utilisation soit de plus en plus limitée [180] en analyse protéomique.

Il existe toutefois plusieurs solutions envisageables pour améliorer les résultats obtenus par l'approche PMF. Il est par exemple possible de remplacer les instruments MALDI-TOF par des appareils plus performants tels que des instruments hybrides Q-TOF ou des analyseurs à ultra-haute résolution. Autrement, il est possible de réaliser une protéolyse partielle des histones par différents moyens dont nous discuterons par la suite. Cette protéolyse partielle pourra générer des peptides plus longs et potentiellement plus spécifiques réduisant ainsi la complexité du mélange et les peptides de même masse. Enfin, le couplage hors ligne d'une chaine HPLC avec un spectromètre de masse MALDI-TOF pourrait permettre de préfractionner l'échantillon avant son analyse et de faciliter l'identification de certaines protéines.

IV.2.4.2 Stratégies LC-MS/MS

IV.2.4.2.1 La fragmentation peptidique

Face aux limitations rencontrées lors de l'utilisation des approches PMF, l'utilisation de la spectrométrie de masse en tandem devient indispensable en analyse protéomique. Elle permet de déterminer la séquence en acides aminés de protéines ou de peptides avec une grande spécificité. La spectrométrie de masse en tandem implique deux étapes d'analyse en masse : la mesure de masse ainsi que la sélection de l'ion précurseur, puis la mesure de masse des ions fragments (figure 46).

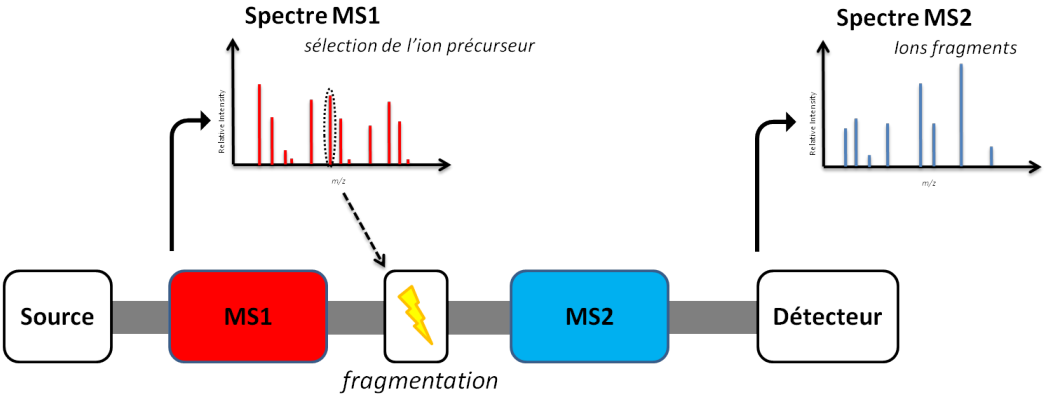


Figure 46 : représentation schématique d'une expérience de spectrométrie de masse en tandem.

La fragmentation peut faire intervenir divers mécanismes en phase gazeuse qui donneront naissance à différents types d'ions fragments. Il existe actuellement deux modes de fragmentation utilisés en protéomique qui sont disponibles sur les instruments commerciaux : la fragmentation induite par collision (CID) et la dissociation par transfert/capture d'électrons (ETD/ECD).

- La fragmentation induite par collision (CID)

La fragmentation induite par collision (*Collision-Induced Dissociation*, CID) est un mode de fragmentation impliquant la collision à basse énergie des ions en phase gazeuse avec des molécules de gaz inerte, généralement l'Argon. La fragmentation peptidique ne se fait pas de manière anarchique mais répond à des règles bien précises établies par Roepstorff *et al.*[181] en 1984 puis par Biemann[182] en 1990.

D'après cette nomenclature, les ions dont la charge positive est portée par la partie N-ter du peptide fragmenté appartiennent aux séries a, b et c, tandis que ceux dont la charge positive est portée par la partie C-ter appartiennent aux séries x, y et z (figure 47).

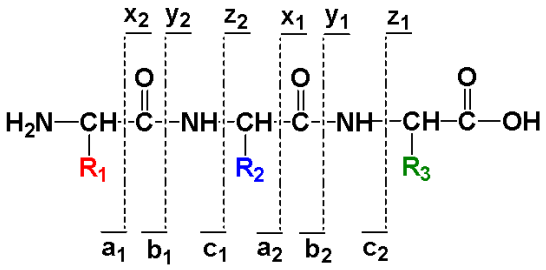


Figure 47 : nomenclature des différentes séries d'ions fragments selon Biemann. R1, R2 et R3 correspondent à des chaines latérales d'acides aminés. D'après http://www.chem.purdue.edu/hilkka/pepseq2.gif.

Les séries d'ions b et y sont généralement majoritaires puisque ces ions proviennent de la rupture de la liaison peptidique qui est une des moins stabilisées. Au sein d'une même série d'ions, les écarts de masse entre ions consécutifs permettent de déterminer la séquence en acides aminés du peptide fragmenté (tableau 7).

Tableau 7 : masses monoisotopiques des différents résidus d'acides aminés

Acides aminés	Code à une lettre	Masse monoisotopique (Da)
Glycine	G	57,02147
Alanine	A	71,03712
Sérine	S	87,03203
Proline	P	97,05277
Valine	V	99,06842
Thréonine	T	101,04768
Cystéine	C	103,00919
Isoleucine	I	113,08407
Leucine	L	113,08407
Asparagine	N	114,04293
Aspartate	D	115,02695

Glutamine	Q	128,05858
Lysine	K	128,09497
Glutamate	E	129,04260
Méthionine	M	131,04049
Histidine	H	137,05891
Phénylalanine	F	147.06842
Arginine	R	156,10112
Tyrosine	Y	163,06333
Tryptophane	W	186,07932

La fragmentation CID des peptides permettra donc de déterminer les séquences en acides aminés des histones et de leurs variants afin de les identifier avec certitude, mais elle permettra également d'étudier leurs modifications post-traductionnelles. En effet, toute modification chimique covalente d'un acide aminé conduit à une variation de sa masse[183] et est donc indentifiable à partir de son incrément de masse mesuré sur les spectres MS/MS (tableau 8). Dans certains cas, des peptides peuvent porter différentes modifications post-traductionnelles isobares, comme c'est le cas avec l'acétylation et la tri-méthylation. Pour distinguer ces deux modifications, il sera nécessaire de recourir à la spectrométrie de masse à ultra-haute résolution.

Tableau 8 : masses monoisotopiques des différentes modifications post-traductionnelles des histones.

Modification post-traductionnelle	**Incrément de masse monoisotopique (Da)**
Mono-méthylation	14,01565
Di-méthylation	28,03130
Tri-méthylation	42,04695
Acétylation	42,01056
Crotonylation	68,02300
Phosphorylation	79,96633
Ubiquitinylation	114,04292
Biotinylation	226,07760

ADP-ribosylation	541,06111

La fragmentation CID présente également un autre inconvénient. En effet, l'énergie transmise à une molécule lors de la fragmentation CID reste limitée, ce qui restreint son degré potentiel de fragmentation. Pour éviter ces inconvénients, d'autres modes de fragmentation ont été développés.

- La dissociation par capture/transfert d'électrons (ECD/ETD)

La dissociation par capture d'électron (*Electron Capture Dissociation*, ECD) utilise non pas des molécules de gaz, mais un faisceau d'électrons lents pour fragmenter les molécules[184]. Elle est basée sur la capture d'un électron par un ion positif multichargé. L'exothermicité de la réaction d'attachement de l'électron au polycation fournit l'énergie nécessaire à la fragmentation. Néanmoins, les contraintes techniques nécessaires à la mise en œuvre de ce type de fragmentation (vide poussé, champ statique) la réserve exclusivement à certains instruments (FT-ICR). La dissociation par transfert d'électrons (*Electron Transfert Dissociation*, ETD) est un mode de fragmentation des peptides et des protéines qui a donc été développé par la suite[185] pour s'affranchir de ces contraintes. Son principe repose sur le transfert d'un électron de faible énergie depuis un anion en phase gazeuse (généralement le fluoranthène) sur une molécule protonée *via* des réactions ion-ion. Ce transfert d'électron entraîne la conversion de la molécule protonée ayant un nombre d'électrons pair en un cation radical qui pourra se dissocier de différentes manières[186], indépendamment de la séquence en acides aminés. Généralement, l'information fournie par ce type de fragmentation est plus riche que par CID puisqu'on retrouve de très nombreux fragments, et notamment la présence de séries d'ions c et z. Ce gain d'information entraîne inévitablement une difficulté supplémentaire lors de l'interprétation des spectres ainsi qu'une perte non négligeable de sensibilité, obligeant à travailler avec des quantités d'échantillons plus importantes. La fragmentation ETD a la particularité d'être plus efficace sur les peptides de grande taille et de préserver les modifications post-traductionnelles. Depuis peu, elle est également appliquée à l'étude des protéines entières[187].

IV.2.4.2.2 Stratégie « bottom-up »

La stratégie la plus courante en analyse protéomique est la stratégie dite « *bottom-up* »[153]. Cette stratégie est basée sur le clivage enzymatique des protéines en peptides généralement de 8 à 25 acides aminés qui sont ensuite séparés par RP-HPLC (colonne C_{18}) couplée en ligne à un spectromètre de masse utilisant la fragmentation CID. La trypsine est l'endoprotéase la plus utilisée pour ce type d'approche. Elle garantit la spécificité et l'efficacité du clivage en C-ter des résidus arginine et lysine. L'avantage principal de cette méthode est de travailler sur des peptides de petite taille ce qui offre une meilleure exactitude sur la mesure de masse et un séquençage plus efficace. Cela permet d'identifier avec davantage de certitude les protéines et leurs modifications post-traductionnelles en comparant les séquences des peptides expérimentaux avec celles générées *in silico* par l'intermédiaire des banques de séquences protéiques en ligne.

Cependant, cette stratégie qui utilise classiquement la trypsine pour le clivage enzymatique n'est pas réellement appropriée pour l'étude des histones. En effet, les résidus lysine et arginine sont très abondants au sein des histones, particulièrement au niveau de leurs extrémités N-ter. Une trypsinolyse générera donc de nombreux peptides de très petite taille (< 7 acides aminés) qui conduiront à une perte de l'information combinatoire du code histone. Les peptides étant très courts, ils seront rarement protéotypiques et les différents types d'histones étant généralement digérés en mélange, il sera impossible de savoir quel peptide provient de quel variant d'histones. La stœchiométrie des différentes formes d'histones sera donc totalement perdue et l'information sera morcelée à l'image d'un puzzle. De plus, l'attribution des spectres MS/MS aux différentes protéoformes pourra parfois être discutable du fait de l'absence de mesure de masse de la protéine entière avant protéolyse. Pour parer à cela, il existe des solutions visant à augmenter la taille moyenne des peptides protéolytiques. La première consiste à utiliser une autre endoprotéase, notamment l'endoprotéase Arg-C, qui clive spécifiquement en C-ter des résidus arginine. Cependant cette protéase ne possède pas la robustesse et la reproductibilité de la trypsine, qui reste le meilleur choix. L'autre solution consiste à dériver chimiquement toutes les lysines libres et chargées positivement afin de neutraliser leur charge comme lorsqu'elles sont acétylées et d'empêcher ainsi la trypsine de cliver la laison

adjacente à ces résidus. Cette solution revient donc à mimer une protéolyse par l'endoprotéase Arg-C tout en utilisant la trypsine. Cette dérivation chimique utilise généralement l'anhydride acétique ou l'anhydride propionique, qui présentent également l'avantage d'augmenter l'hydrophobicité des peptides et d'améliorer leur séparation chromatographique[188].

L'approche « *bottom-up* » présente donc des avantages sérieux en termes de séparation et de résolution chromatographique des peptides ainsi que de qualité des spectres MS/MS acquis. C'est une stratégie relativement simple à mettre en place et qui est quasi-universelle. Elle peut s'avérer utile comme première approche pour caractériser les histones contenues dans un échantillon ainsi que leurs modifications post-traductionnelles, ou alors dans le cas d'un sous-type d'histone préalablement purifié par SDS-PAGE ou par HPLC. Lorsqu'il s'agit de caractériser finement le code histone, elle trouve rapidement ses limites.

IV.2.4.2.3 Stratégie « top-down »

La stratégie « top-down » utilise les protéines entières comme matériel de départ et non les peptides protéolytiques. Les protéines intactes peuvent être introduites soit par infusion directe soit en sortie de colonne chromatographique. Les histones étant de petites protéines, cette approche se prête particulièrement bien à leur caractérisation par spectrométrie de masse. Des protéines entières allant jusqu'à 200 kDa ont même pu être caractérisées par l'approche « *top-down* »[189]. La préparation des échantillons pour l'analyse « *top-down* » est plus rapide que dans le cas d'une analyse « *bottom-up* » puisqu'elle ne nécessite pas d'étape de protéolyse enzymatique. Elle permet une analyse plus complète des séquences protéiques et des profils de modifications post-traductionnelles. A côté de cela, cette approche présente elle aussi quelques inconvénients. Le premier concerne l'efficacité d'ionisation des protéines intactes en ESI, qui est moindre que celle des peptides. Ensuite, elle requière une instrumentation plus élaborée ainsi que des logiciels adaptés pour analyser les données. Les spectromètres de masse compatibles avec ce type d'approches utilisent la fragmentation ECD ou ETD. Ils doivent être particulièrement performants en termes de résolution et d'exactitude sur la mesure de masse.

Les spectres MS/MS générés par les approches « *top-down* » sont extrêmement complexes. Les ions fragments sont très nombreux et présents à différents états de charge. Il est donc difficile d'identifier et de localiser précisément les modifications post-traductionnelles présentes sur les histones. Mais malgré les difficultés rencontrées, cette approche a fait ses preuves et a conduit à la publication de nombreuses études. La première analyse des modifications post-traductionnelles des histones par une approche « *top-down* » a été publiée par Banks *et al.*[190] en 2001. Cette étude portait sur la déphosphorylation des isoformes de H1 sous l'effet de la dexaméthasone. Après avoir été pré-purifiées par HPLC, les fractions ont été directement injectées dans un spectromètre de masse de type Q-TOF.

Ce n'est qu'avec les travaux de Kelleher et de son équipe que l'approche « *top-down* » a émergé en protéomique des histones. En 2006, ils ont publié une série d'articles sur la caractérisation de chacun des sous-types d'histone de cœur par une approche « *top-down* ». En ayant recours à l'infusion directe d'histones pré-purifiées par RP-HPLC et à la fragmentation ECD, ils sont parvenus à révéler la diversité des principaux variants des histones H2A, H2B, H3 et H4[191–194].

Actuellement, la stratégie « *top-down* » repose sur une séparation efficace des histones intactes avant leur analyse MS et MS/MS. Le couplage classique LC-MS/MS reste le choix le plus judicieux puisqu'il permet d'effectuer à haut débit des analyses de manière automatique et reproductible. Deux types de chromatographie liquide ont été utilisés avec succès : la RP-HPLC et l'HILIC. Malgré des résultats satisfaisants obtenus en chromatographie HILIC évoqués précédemment, la phase stationnaire de choix reste la silice greffée C_{18}. Plus récemment, un système complexe de chromatographie bidimensionnelle a été proposé par Tian *et al.*[195] pour caractériser les histones et leurs modifications post-traductionnelles. Parallèlement à cela, deux équipes ont publié leurs travaux utilisant l'UPLC. La première équipe a réalisé une analyse « top-down » en utilisant l'UPLC puis une fragmentation des protéines intactes par ETD[196]. La seconde a également utilisé l'UPLC toujours sur des histones intactes, mais a ensuite choisi de recourir à la fragmentation CID. Ce choix nécessite une étape d'analyse supplémentaire sur les peptides afin d'annoter correctement les différentes modifications post-traductionnelles identifiées sur les histones intactes[176].

Cette stratégie semble donc parfaitement adaptée aux histones en préservant l'information combinatoire, même si la distribution du signal entre les très nombreux états de charge entraîne une perte de sensibilité et que le comportement chromatographique des protéines intactes n'est pas aussi reproductible que celui des peptides[197].

IV.2.4.2.4 Stratégie « middle-down »

Les limites respectives des approches « *bottom-up* » et « *top-down* » ont inévitablement poussé les spécialistes de la protéomique à développer une approche intermédiaire qui bénéficierait à la fois de la sensibilité et de la résolution de l'approche « *bottom-up* » tout en préservant l'intégralité de l'information combinatoire fournie par l'approche « *top-down* »[198]. Cette approche baptisée naturellement « *middle-down* » est basée sur l'étude des peptides, mais contrairement à l'approche « *bottom-up* » elle vise à générer des peptides de taille plus importante (40 à 50 acides aminés). Elle se base sur le fait que la majorité des modifications post-traductionnelles des histones de cœur surviennent en N-ter, sur les 50 premiers acides aminés. Les histones peuvent être clivées en gros polypeptides par des endoprotéases autres que la trypsine ayant des spécificités différentes comme par exemple l'endoprotéase Asp-N qui coupe spécifiquement en N-ter des résidus aspartate ou encore l'endoprotéase Glu-C qui coupe spécifiquement en C-ter des résidus glutamate et aspartate dans une moindre mesure. L'analyse LC-MS/MS de ces polypeptides peut faire appel à la fragmentation CID ou ECD/ETD.

La première étude réalisée sur les histones utilisant cette approche a été publiée par Taverna *et al.*[199] en 2007. Par la suite, Young *et al.* ont réussi, en utilisant l'approche « *middle-down* », à caractériser le plus grand nombre de combinaisons de modifications post-traductionnelles sur les histones H3.2 et H4 en une seule expérience[200]. De manière générale, cette stratégie représente un compromis intéressant entre les deux approches précédentes. La séparation chromatographique des polypeptides reste moins délicate que celle des protéines intactes. L'HILIC s'est d'ailleurs révélée être très utile dans ce type d'approches puisqu'elle permet par exemple de séparer les peptides isobares acétylés et tri-méthylés d'après leur temps de rétention, les peptides tri-méthylés étant

davantage retenus sur la colonne[201]. Garcia et ses collaborateurs ont en parallèle démontré que les résines mixtes échange d'ions/HILIC offraient la meilleure résolution pour l'étude des extrémités N-ter des histones[202]. Au final, la stratégie « *middle-down* » se présente de plus en plus comme étant la stratégie la plus performante pour l'étude des histones et de leurs modifications post-traductionnelles et continue de faire l'objet d'améliorations constantes[203]. Cependant, elle reste relativement lourde a mettre en place et nécessite d'avoir a disposition les appareils et les logiciels de traitement de données adéquats.

Il semble donc que la stratégie parfaite n'existe pas et que la meilleure solution soit probablement de les considérer comme complémentaires (figure 48). Concernant l'analyse des données, le principe reste le même pour les trois approches. Il consiste soit à comparer les spectres MS/MS expérimentaux obtenus par fragmentation avec des spectres MS/MS théoriques, ce que l'on appelle l'empreinte de fragmentation peptidique (*Peptide Fragment Fingerprinting*, PFF), soit à déduire les séquences d'acides aminés directement des écarts de masse entre pics successifs d'une même série d'ions, ce que l'on appelle le séquençage *de novo*. Ces approches trouvent donc leurs limites lorsqu'il s'agit non plus de caractériser mais de quantifier les variants d'histones et leurs modifications.

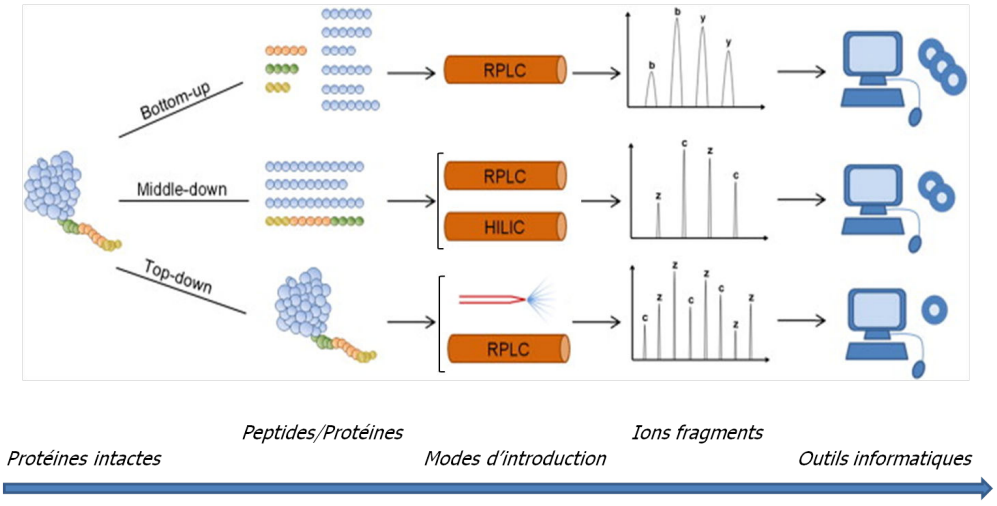


Figure 48 : comparaison des trois stratégies d'analyse des histones par spectrométrie de masse. Adapté de[204].

IV.2.4.2.5 Approches quantitatives

Nous venons de passer en revue les méthodes classiquement utilisées pour l'analyse du code histone par spectrométrie de masse. Cependant, ces méthodes fournissent principalement une information qualitative et non quantitative. En effet, la diversité des protéoformes d'histones nous prive de l'utilisation d'un standard interne. Dans le cadre d'une étude toxicologique de l'effet d'un xénobiotique sur le code histone, l'abondance relative des différents variants d'histones et de leurs modifications post-traductionnelles est capitale. La régulation épigénétique pourra jouer sur la nature ou le degré de modification des histones ainsi que sur la quantité de variants incorporés. C'est donc l'abondance relative des variants d'histones et la dynamique de leurs modifications post-traductionnelles qu'il faut parvenir à appréhender. Pour cela, certaines méthodes quantitatives en protéomique permettront de comparer le contenu en histones d'échantillons provenant de cellules saines et de cellules exposées à un toxique. Elles vont du marquage métabolique de protéines par des isotopes stables en culture cellulaire[205] (*Stable Isotope Labelling with Amino acids in Cell culture*, SILAC) à la dérivation chimique par des isotope stables en passant par des méthodes sans marquage dites « *label-free* »[206]. Il existe donc des méthodes avec marquage qui reposent sur la mesure du rapport des intensités entre la même espèce moléculaire marquée différentiellement par un isotope lourd ou un isotope léger. Les méthodes sans marquage utilisent simplement la comparaison de l'intensité des signaux correspondant aux différents ions, ou encore le comptage du nombre de spectres acquis pour chaque espèce qui est proportionnel à son abondance. Plusieurs revues récentes décrivent l'utilisation de la protéomique quantitative pour l'étude du code histone[207,208]. Une des premières études quantitative des histones par spectrométrie de masse a été publiée par Smith *et al.* en 2003[209]. Ils y présentent la quantification des niveaux d'acétylation de l'histone H4 en utilisant une dérivation chimique par l'anhydride acétique deutéré. Un marquage semblable par l'anhydride propionique a par la suite été proposé par Plazas-Mayorca et ses collaborateurs[210]. Comme nous l'avons évoqué précédemment, l'anhydride propionique réagit avec les lysines libres chargées positivement et neutralise leur charge, empêchant ainsi la trypsine de cliver après ces résidus. En utilisant un anhydride propionique léger d_0 et un anhydride

propionique lourd d_{10} il est possible de marquer différentiellement deux échantillons provenant de deux conditions différentes et de réaliser une quantification relative. C'est d'ailleurs aujourd'hui la méthode la plus utilisée pour la quantification relative des histones.

La quantification par SILAC est également utilisée dans de nombreuses études sur la dynamique de la chromatine. Le marquage *in vivo* des protéines à l'échelle métabolique se fait par incorporation d'acides aminés marqués. Il permet de combiner les différentes conditions expérimentales à un stade très avancé et d'éviter au maximum les biais introduits lors des différentes étapes de préparation des échantillons. La méthode SILAC a donc naturellement été utilisée en combinaison avec la spectrométrie de masse pour suivre la dynamique des modifications post-traductionnelles des histones au cours du cycle cellulaire[211]. D'autres études ont également permis de suivre simultanément l'abondance relative de plusieurs modifications post-traductionnelles, conservant ainsi l'information combinatoire. Cette approche s'est révélée être très intéressante pour étudier la dynamique des histones et de leurs modifications post-traductionnelles à partir de cultures cellulaires[212]. Elle reste cependant très chronophage, coûteuse et réservée à l'étude des histones extraites de cellules en culture.

Une autre méthode de quantification faisant appel cette fois au marquage des peptides et à l'analyse LC-MS/MS a été utilisée en biologie de la chromatine[213]. Il s'agit du marquage isobare pour la quantification relative et absolue[214] (*isobaric Tag for Relative and Absolute Quantitation*, iTRAQ®). Elle consiste à fixer en N-ter des peptides différents réactifs isobares composés d'un rapporteur et d'un équilibreur. Les masses de chacune des parties sont différentes pour chaque réactif. Les peptides marqués par un réactif iTRAQ® seront donc détectés sur les spectres MS1 avec un incrément de masse qui leur est propre. La fragmentation CID des peptides marqués entraîne la perte des fragments équilibreurs neutres et donc la libération d'ions rapporteurs de masses différentes utilisés pour la quantification. Avec les kits iTRAQ® modernes, il est possible de comparer jusqu'à 8 états cellulaires différents. Cependant, les études publiées utilisant cette approche n'ont pas permis de fournir des informations précises sur l'abondance des différentes modifications post-traductionnelles. L'appellation iTRAQ® est un nom

déposé par la société AB SCIEX, mais il existe des kits équivalents chez d'autres fournisseurs, comme le TMT chez Thermo Scientific par exemple.

Les approches « *label-free* » représentent une alternative intéressante aux approches avec marquage. En utilisant les ratios des abondances relatives des différents ions fragments sur les spectres MS/MS, elles permettent de réaliser une quantification relative des différentes formes d'histones, notamment des espèces isobares qui co-éluent mais qui ne présenteront pas le même profil d'ions fragments. Ces approches nécessitent néanmoins de recourir à des instruments à ultra-haute résolution. La quantification de peptides spécifiques pourra également être réalisée par des approches ciblées en suivant des transitions ions précurseurs / ions fragments sur des appareils de type triple quadripôle. Ces dernières approches restent trop ciblées pour être utilisées dans le cas d'une analyse complète du code histone.

En résumé, toutes les méthodes de quantification avec (figure 49) ou sans marquage possèdent des avantages, mais aussi des inconvénients lorsqu'il s'agit de l'étude exploratoire du code histone dans un contexte de perturbation environnementale.

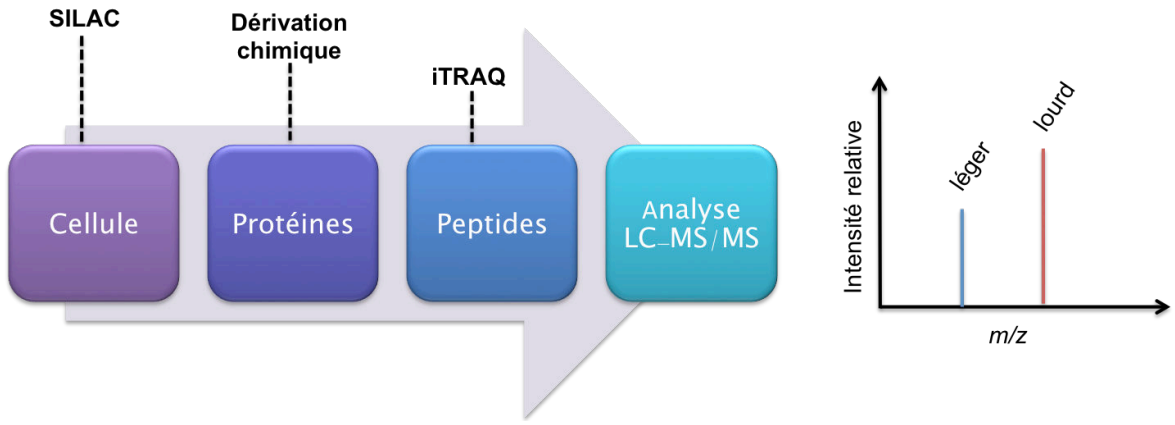


Figure 49 : stratégies de marquage utilisées pour la quantification des histones par spectrométrie de masse.

IV.2.4.3 Conclusion

Dans le cadre d'une étude toxicologique de l'impact d'un xénobiotique sur la régulation épigénétique, notre objectif n'est pas de caractériser finement la nature et la localisation de chacune des modifications post-traductionnelles, ni de quantifier quelques histones spécifiques parmi la multitude d'espèces présentes. Notre principal objectif est de déterminer de la manière la plus rapide et la plus simple possible si un xénobiotique perturbe l'incorporation des variants histones ou la dynamique de leurs modifications post-traductionnelles. La méthode employée devra fournir rapidement une réponse à la question suivante : ce xénobiotique induit-il une perturbation significative du code histone ? Elle devra donc permettre de comparer de manière globale le contenu en histones d'échantillons témoins *versus* celui d'échantillons exposés, et ce en une seule expérience. Cette comparaison simultanée de l'abondance de l'ensemble des espèces détectées permettrait ainsi de conserver l'aspect combinatoire et d'extraire un ensemble de marqueurs histoniques caractéristiques d'un état cellulaire, autrement dit d'une exposition. A notre connaissance, au moment où nous avons entrepris ces travaux, il n'existait aucune méthode publiée répondant à ces critères et adaptée aux histones. C'est pourquoi nous avons travaillé à développer une méthode complémentaire qui permette de visualiser l'ensemble des espèces simultanément afin d'étudier le code histone en tant qu'unité et non en tant que somme de ses constituants.

Partie 2

L'APPROCHE HISTONOMIQUE GLOBALE

Chapitre I

Principe général

Chapitre II

Obtention des histones à partir de cellules humaines

Chapitre III

Profilage des histones intactes par LC-MS

Chapitre IV

Prétraitement des données

Chapitre V

Approches statistiques multivariées pour l'analyse des données

Chapitre VI

Interprétation et validation des résultats

I. Principe général de l'approche histonomique

Le but du chapitre précédent était de donner une vue d'ensemble des différentes stratégies basées sur l'utilisation de la spectrométrie de masse pour déchiffrer le code histone quel que soit le contexte biologique. Nous avons vu pourquoi le recours à la spectrométrie de masse était indispensable et comment les différentes stratégies permettaient d'obtenir une information davantage qualitative ou *a contrario* quantitative. Toutes les approches présentées continuent d'être très utilisées pour l'étude du code histone, et le but de notre travail n'est surtout pas de tenter de prouver qu'elles présentent plus d'inconvénients que d'avantages. Au contraire, nous avons tenté de mettre en avant la puissance de ces approches et le niveau de détail auquel elles permettent d'accéder. Cependant, notre étude se place dans un contexte très particulier. Ce travail de thèse s'est intégré au sein d'un projet ANR baptisé PLACENTOX, dont l'objectif était de mieux comprendre les mécanismes moléculaires et épigénétiques de toxicité placentaire induits par une exposition de la femme enceinte à des polluants environnementaux. La tâche qui nous a été confiée consistait à associer un profil de modifications d'histones à l'exposition à un polluant donné, afin de pouvoir classer les patients et extraire des marqueurs spécifiques d'exposition. La nécessité de discriminer les patients sur la base de leur code histone nous a amené à développer un outil de screening puissant et rapide nous permettant d'obtenir suffisamment d'informations sans avoir accès directement à la localisation précise des modifications post-traductionnelles. En effet, avant de mettre en œuvre une stratégie de caractérisation fine de chacune des modifications, il nous est apparu nécessaire de déterminer si le polluant en question était à l'origine d'une perturbation du profil d'histones. Les questions auxquelles notre approche devait pouvoir répondre étaient les suivantes : les profils d'histones des patients sains et des patients exposés sont-ils différents ? Quelles sont les histones impactées ?

Nous avons donc développé un outil de classification des échantillons sur la base de leur profil d'histones. Pour développer et valider cette approche que nous avons baptisée approche histonomique globale, nous avons choisi d'utiliser dans un premier temps des mélanges standards d'histones commerciales ainsi que des histones extraites de cellules en culture non exposées. Nous avons tenté

d'améliorer les méthodes d'extraction des histones décrites dans la littérature et constitué ainsi notre propre protocole. Nous avons ensuite choisi de travailler sur des histones intactes et non sur des peptides, afin de réduire la complexité des mélanges, de minimiser les étapes de préparation d'échantillon et surtout de préserver l'information combinatoire. L'UPLC couplée à un spectromètre de masse de type ESI-QTOF a été choisie pour réaliser le profilage des histones intactes, puis une stratégie de prétraitement et d'analyse multivariée des données a été mise en place, s'inspirant de ce qui se fait en analyse métabolomique. Nous présenterons dans ce chapitre la mise au point de l'ensemble des étapes de l'approche histonomique (figure 50).

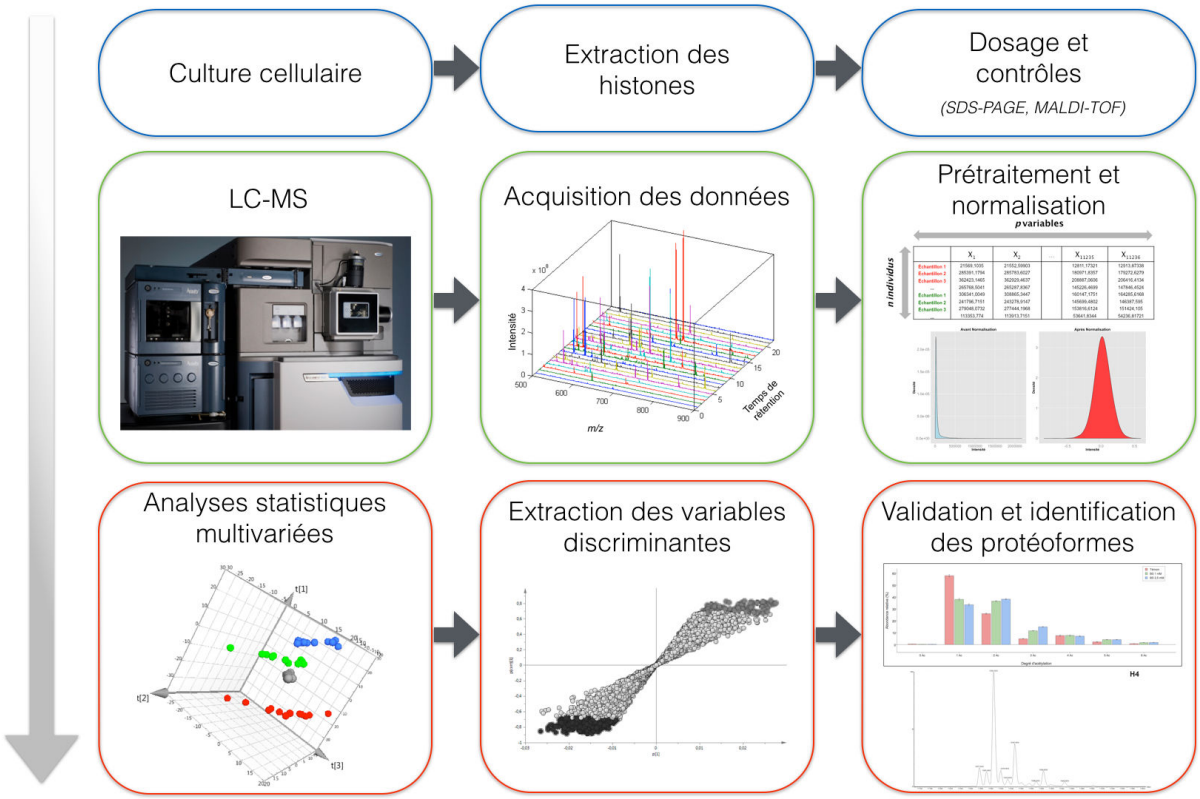


Figure 50 : étapes séquentielles de l'approche histonomique globale.

II. Obtention des histones à partir de cellules humaines

Le but final de ce projet ANR étant de travailler sur des primo-cultures de cellules trophoblastiques directement issues de placentas humains, nous avons choisi de mettre au point et de valider l'ensemble des étapes constitutives de notre approche sur du matériel biologique moins précieux mais tout aussi pertinent d'un point de vue biologique : des histones extraites de cellules placentaires BeWo.

II.1 Culture cellulaire

Afin de mimer *in vitro une* exposition des cellules trophoblastiques humaines à divers xénobiotiques, la lignée cellulaire BeWo[215] s'est révélée être le meilleur choix. En effet, les cellules BeWo, issues de choriocarcinome humain, sont un bon modèle de cytotrophoblastes. Elles sont faciles d'accès et relativement simples à cultiver. Elles possèdent des fonctions endocrines similaire aux syncytiotrophoblastes et expriment les mêmes protéines de transport[216] et les mêmes enzymes du métabolisme[217] que les trophoblastes du troisième trimestre[218]. Le clone b30 que nous avons utilisé au cours de ces travaux présente la capacité de former une monocouche de cellules confluentes en une période de temps assez courte. Leur principale différence avec les trophoblastes *in vivo* réside dans le fait qu'elles ne se différencient que très peu spontanément et sont présentes majoritairement sous forme de cytotrophoblastes non syncytialisés. Dans le cadre de ce projet, les cellules BeWo clone b30 nous ont été fournies gracieusement par l'équipe du Dr. Danièle Evain-Brion (ex- UMR-S767) sous forme d'un cryotube de cellules congelées. Les cellules avaient auparavant été obtenues auprès du laboratoire du Dr. Alan Schwartz de l'Université de Washington aux Etats-Unis. Les cellules BeWo commercialement disponibles auprès de l'ATCC (*American Type Culture Collection*) ne possèdent pas exactement les mêmes propriétés que le clone b30.

Le cryotube fourni contenait environ 3 millions de cellules congelées dans un milieu composé de 90% de milieu de culture F-12K (modification de Kaighn) et de 10% de diméthylsulfoxyde (DMSO). La décongélation rapide s'est faite selon le protocole détaillé dans la partie expérimentale. Les cellules BeWo ont ensuite été

mises en culture dans du milieu F-12K supplémenté avec 10% de sérum de veau fœtal (SVF) afin de garantir la présence des nutriments nécessaires à leur croissance. De plus, afin d'éviter les contaminations, le milieu de culture a systématiquement été supplémenté avec deux antibiotiques : la pénicilline à 50 $UI.mL^{-1}$ et la streptomycine à 50 $UI.mL^{-1}$. Dans la suite du manuscrit, le milieu complet fera référence au milieu F-12K contenant 10% de SVF ainsi que les deux antibiotiques à 50 $UI.mL^{-1}$ chacun. Les cellules ont été maintenues dans des conditions standard de culture, à savoir dans un incubateur avec une atmosphère à 37°C et 5% CO_2.

Dans les conditions de culture que nous venons de décrire, nous sommes parvenus à obtenir une flasque à confluence (figure 51) en 4 à 5 jours seulement.

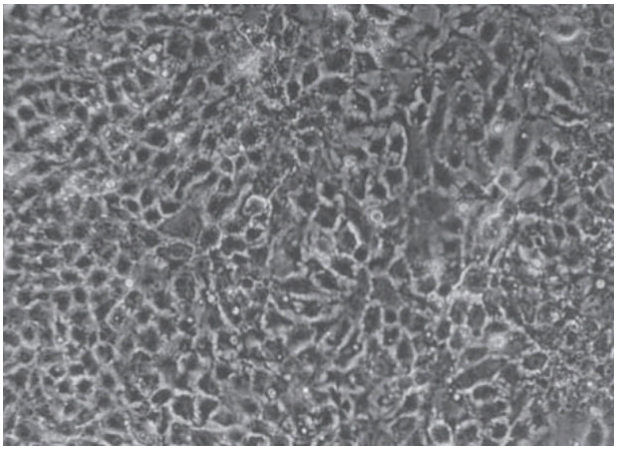


Figure 51 : photographie d'une monocouche de cellules adhérentes à confluence BeWo clone b30 cultivées dans une flasque T75 et observées au microscope (grossissement x200).

Une fois les cellules arrivées à confluence, elles ont été décollées et individualisées par digestion trypsique des protéines d'adhésion, puis comptées sur une cellule de numération Neubauer. D'après ce comptage, une flasque T75 à confluence contenait en moyenne entre 5 et 6 millions de cellules. A chaque passage (trypsination) chacune des flasques était répartie dans 5 nouvelles flasques (cf. protocole en partie expérimentale). Le nombre de passages réalisés est une information importante, puisqu'une lignée cellulaire ne peut supporter qu'un nombre limité de passages avant de présenter un phénotype aberrant ou de mourir. Le nombre maximal de passage dépend de la lignée cellulaire, et dans le cas des cellules BeWo nous l'avons fixé à 23.

Pour obtenir des culots cellulaires en vue de réaliser une extraction des histones dont nous parlerons lors du chapitre suivant, le protocole appliqué a été le même que lors des passages, à la seule différence près que la suspension cellulaire issue d'une flasque n'a pas été répartie entre plusieurs nouvelles flasques mais intégralement transférée dans un tube Falcon avant centrifugation et lavage par du tampon phosphate salin froid (*Phosphate Buffered Saline*, PBS). Après élimination du surnageant, les culots secs obtenus ont été conservés au congélateur à -80°C jusqu'à utilisation.

II.2 Extraction des histones

L'extraction des histones à partir de cellules ou de tissus repose principalement sur leur caractère basique et donc sur leur solubilité en milieu acide. Les méthodes couramment employées, comme celle décrite par Shechter *et al.*[219] en 2007, utilisent donc un acide dilué (acide sulfurique H_2SO_4 ou acide chlorhydrique HCl) pour extraire les histones. La quasi totalité de ces méthodes réalise en premier lieu une lyse des membranes cytoplasmiques afin de purifier les noyaux. L'extraction par un acide dilué se fait alors directement sur les noyaux. Cependant, même si la plupart des protéines nucléaires et des acides nucléiques précipitent dans ces conditions[220], il peut subsister après extraction quelques protéines autres que les histones sous forme soluble dans le surnageant de précipitation. Ces contaminants dont les propriétés physico-chimiques sont très proches de celles des histones seront difficiles à éliminer et pourront compliquer davantage l'analyse des histones par la suite. Il est en effet fondamental d'obtenir un extrait d'histones aussi pur que possible pour éviter toute ambiguïté d'identification.

Pour cela, nous avons testé plusieurs conditions et apporté quelques modifications au protocole classique de Shechter. Tout d'abord, nous avons ajouté une étape de lyse des noyaux par un tampon contenant un détergent non ionique (NP-40) afin d'extraire les histones non plus à partir des noyaux intacts mais à partir de la chromatine isolée, dans le but de réduire au maximum la présence de contaminants protéiques. Nous avons également ajouté quelques étapes de lavage au cours du protocole, et allongé le temps d'incubation avec l'acide sulfurique le faisant passer de 30 minutes à 4 heures. Pour ce qui est de la composition des

tampons, nous avons veillé à utiliser des inhibiteurs de protéases et de phosphatases ainsi que le butyrate de sodium en tant qu'inhibiteur HDAC pour éviter que les enzymes libérées lors de la lyse cellulaire n'agissent pendant toute la durée de l'extraction.

L'extraction acide ayant été décrite comme pouvant altérer certaines modifications post-traductionnelles labiles[221,222], une autre technique plus conservatrice a été proposée par von Holt *et al.*[223]. Cette technique repose sur l'utilisation de tampons salins qui garantit le maintien d'un pH neutre tout au long de l'extraction et préserve ainsi certaines modifications. L'extraction saline permet également d'extraire différentiellement les histones H2A et H2B des histones H3 et H4 en jouant sur la concentration en chlorure de sodium NaCl. Nous avons testé cette extraction afin de comparer la quantité et la pureté des extraits finaux par rapport à l'extraction acide. Cependant, les quantités d'histones obtenues étant plus faibles et le protocole étant plus lourd à mettre en place, nous avons privilégié l'extraction acide.

L'étape suivant l'extraction consiste à faire précipiter les histones en milieu acide. Pour cela, l'acide trichloroacétique (TCA) est un acide extrêmement efficace pour faire précipiter les protéines et il est utilisé de manière universelle. Etrangement, son mécanisme d'action reste peu connu, même si l'agrégation hydrophobe semble être l'hypothèse la plus probable[224]. La précipitation au TCA peut cependant être à l'origine de la formation d'un précipité insoluble qui peut contenir des histones. Nous avons donc également testé plusieurs conditions pour la précipitation au TCA en faisant varier sa concentration finale ou en ajoutant de l'acétone.

Enfin nous avons ajouté plusieurs étapes de lavage à l'acétone afin de se débarrasser des acides qui pourraient gêner l'analyse par spectrométrie de masse.

Nous avons donc comparé plusieurs protocoles différents afin de trouver celui qui permettait d'obtenir le plus facilement possible la plus grande quantité d'histones avec le moins d'impuretés. Nous avons utilisé comme matériel de départ des culots de cellules BeWo non traitées (environ 5 millions de cellules par culot). Le dosage des extraits protéiques totaux permet de se faire une première idée du rendement d'extraction. Pour cela nous avons utilisé la méthode de Bradford[225] qui

repose sur la mesure des absorbances à 595 nm en présence d'un réactif colorimétrique, le bleu de Coomassie, qui se complexifie avec les protéines *via* les acides aminés basiques et aromatiques et change de couleur. Plus l'absorbance est élevée, plus la solution est concentrée en protéines. En comparant une extraction saline avec précipitation au TCA 33% (A), une extraction saline avec une précipitation au TCA 20% (B), une extraction saline avec une précipitation au TCA 10%/acétone 80% (C) et une extraction acide classique par H_2SO_4 0,4 N suivie d'une précipitation au TCA 33% (D), nous avons noté que les quantités obtenues étaient significativement différentes (tableau 9).

Tableau 9 : concentrations moyennes (deux réplicats) en protéines des extraits obtenus par différents protocoles d'extraction. (A) extraction saline et précipitation TCA 33%, (B) extraction saline et précipitation TCA 20%, (C) extraction saline et précipitation TCA 10%/acétone 80% et (D) extraction acide H_2SO_4 0,4 N et précipitation TCA 33%.

Condition	A	B	C	D
Quantité de protéines (µg)	30	26	6	38

Ensuite, les profils électrophorétiques des histones en mélange extraites à partir des différents protocoles ont été obtenus en SDS-PAGE 15% (figure 52). Les gels d'électrophorèse SDS-PAGE 15% ont été coulés selon le protocole détaillé en partie expérimentale. Le bleu de Coomassie colloïdal été utilisé pour effectuer la révélation des bandes.

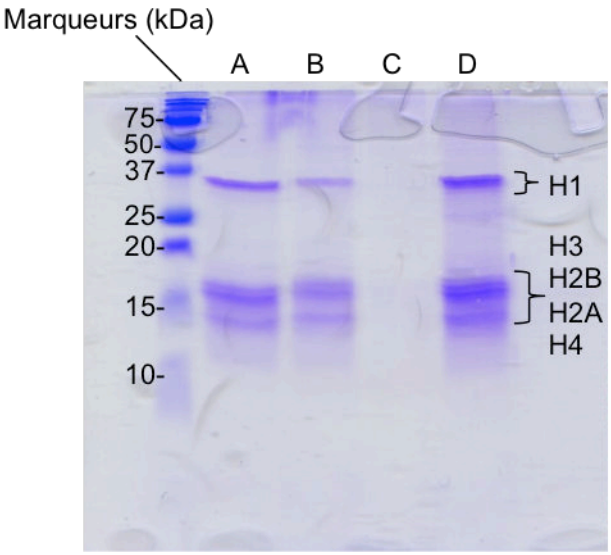


Figure 52 : SDS-PAGE 15% de 4 extraits protéiques obtenus selon les protocoles décrits dans la publication de Shechter avec les modifications citées. Le gel a été étalonné à l'aide de marqueurs de masse moléculaire. A) extraction saline et précipitation TCA 33%, (B) extraction saline et précipitation TCA 20%, (C) extraction saline et précipitation TCA 10%/acétone 80% et (D) extraction acide H_2SO_4 0,4 N et précipitation TCA 33%.

D'après la figure 52 et le tableau 9, la comparaison des échantillons A et D permet de dire que l'extraction la plus efficace semble être l'extraction acide. La quantité de protéines extraites y est plus importante, et l'extrait ne semble pas contenir d'impuretés majeures. De plus, le protocole est plus simple à réaliser et plus robuste. Concernant la précipitation, le meilleur compromis entre l'efficacité de la précipitation et l'absence de précipité insoluble a été obtenu en utilisant une concentration finale en TCA de 25% plutôt que 33%.

Au final, nous avons choisi de retenir l'extraction acide malgré le risque de perdre certaines modifications, notamment les phosphorylations pour certains variants. Le protocole d'extraction acide de départ a donc été amélioré afin de se débarrasser au maximum des impuretés au fil des étapes. Il est détaillé en partie expérimentale.

II.3 Contrôles des extraits histoniques

II.3.1 Dosage des protéines totales

Pour évaluer la quantité d'histones obtenue à partir des culots cellulaires, nous avons systématiquement dosé les extraits finaux ainsi que les fractions conservées au cours de l'extraction correspondant au cytoplasme et au nucléoplasme. Pour réaliser le dosage des protéines, nous nous sommes initialement basés sur la littérature et nous avons utilisé la méthode de Bradford[226]. Cette méthode repose sur l'interaction d'un réactif coloré (bleu de Coomassie G-250 colloïdal) avec les résidus basiques ou aromatiques des protéines. L'intensité de coloration est proportionnelle à la quantité de protéines ce qui a permis de réaliser une gamme standard de calibration en utilisant des quantités définies d'albumine de sérum bovin (BSA). L'intensité de coloration dépend directement de la nature des protéines et de leur richesse en certains acides aminés, or, les histones étant proportionnellement plus riches en acides aminés basiques que la BSA, cela peut aboutir à une surestimation de la quantité de protéines dosée. Après avoir réalisé plusieurs essais, nous avons évalué cette surestimation à environ 30%. Nous avons donc choisi d'utiliser une autre méthode de dosage appelée BCA[227] (*Bicinchoninic acid Assay*). La première étape de cette méthode de dosage colorimétrique est la réduction des ions cuivrique Cu(II) en ion cuivreux Cu(I) par les protéines en milieu alcalin. Ensuite, chaque ion Cu(I) est chélaté par deux molécules d'acide bicinchoninique pour former un complexe dont la coloration évolue vers le pourpre. A 562 nm, l'absorption de ce complexe soluble augmente de façon linéaire avec la concentration en protéines. Contrairement à la méthode de Bradford, la liaison peptidique contribue également à la formation du complexe coloré ce qui minimise la variabilité due aux différences de composition en acides amines. Cette méthode présente également l'avantage d'être moins sensible aux contaminants tels que les détergents.

II.3.2 SDS-PAGE

Une fois le dosage effectué, les extraits protéiques correspondant au cytoplasme, au nucléoplasme et aux histones ont été systématiquement déposés

sur SDS-PAGE 13% afin de contrôler la pureté des extraits histoniques finaux (figure 53). Le détail de la préparation et de la migration des gels est présenté en partie expérimentale.

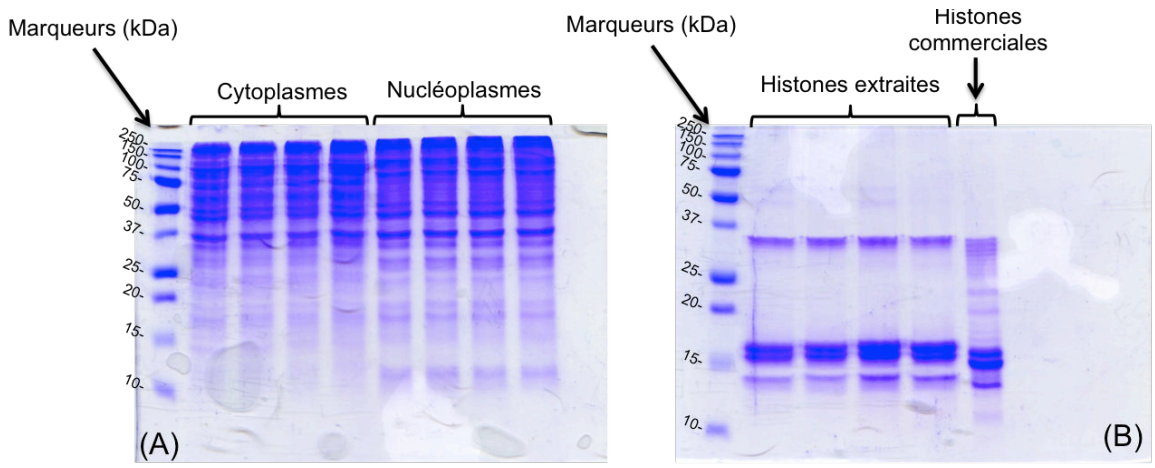


Figure 53 : SDS-PAGE 13% d'extraits protéiques correspondant aux cytoplasmes et aux nucléoplasmes (A) ainsi qu'aux histones (B) extraites à partir de 4 culots identiques de cellules BeWo. Des histones commerciales extraites de thymus de veau ont également été déposées (B) pour comparaison avec les profils obtenus avec notre protocole d'extraction. Environ 10 µg de protéines ont été déposés pour chaque échantillon du gel (A) et environ 5 µg pour ceux du gel (B).

D'après la figue 53, les profils électrophorétiques des 4 réplicats d'extraction semblent parfaitement similaires, ce qui prouve la répétabilité de notre protocole d'extraction. D'autre part seules les bandes correspondant aux différents sous-types d'histones (cf. figure 52) sont visibles, prouvant ainsi l'absence de contaminant majeur et la pureté satisfaisante des échantillons préparés. Ceci est également confirmé en comparant les profils des histones extraites avec celui d'histones commerciales extraites de thymus de veau (Sigma Aldrich, réf. H9250). A quantités équivalentes déposées, le profil des histones commerciales obtenues par extraction acide semble contenir davantage de contaminants que les échantillons extraits par notre protocole. L'intérêt du SDS-PAGE reste cependant limité dans le cas de l'étude des histones en mélange. Il faut également garder en tête que les protéines basiques que sont les histones présenteront généralement une surcoloration au bleu de Coomassie.

II.3.3 Contrôle en MALDI-TOF

Afin de s'assurer réellement de la présence majoritaire en histones de nos échantillons avant de les introduire sur le système LC-MS, une analyse MALDI des extraits histoniques a été réalisée. Les échantillons ont été déposés sur une plaque MALDI après avoir été mélangés à une solution saturée de matrice (acide α-cyano-4-hydroxycinnamique à 5 mg/mL, cf. partie expérimentale). L'instrument dont nous disposons au laboratoire, un MALDI-TOF Voyager-DE^{TM} PRO (de la société Applied Biosystems), a été utilisé en mode linéaire positif.

Le spectre présenté à la figure 54 montre que les espèces majoritairement présentes correspondent aux histones de cœur H2A, H2B, H3 et H4 simplement chargées. Les mêmes espèces doublement chargées sont également présentes sur le spectre. Il est évident que ce type d'appareil ne possède pas la résolution suffisante pour distinguer des espèces de masses très proches, particulièrement en mode linéaire (de l'ordre de 2 000). La complexité du mélange d'histones extraites n'est en rien résolue par cette analyse MALDI-TOF, mais elle permet simplement de s'assurer de l'absence de contaminants majeurs.

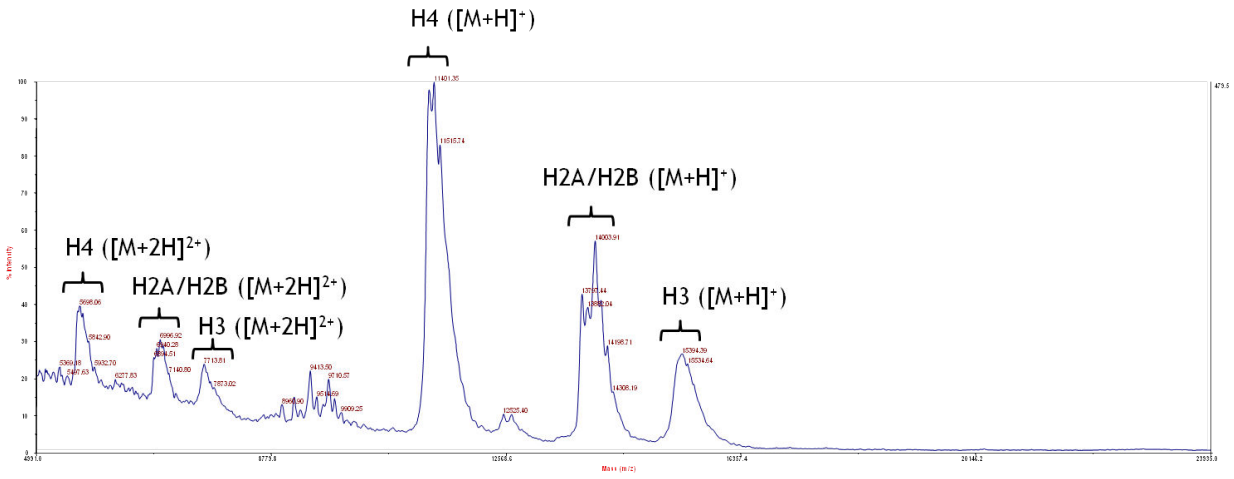


Figure 54 : spectre MALDI-TOF d'histones en mélange extraites à partir d'un culot de cellule BeWo selon notre protocole d'extraction.

III. Profilage des histones intactes par UPLC-MS

Le profilage des histones intactes par chromatographie liquide couplée à la spectrométrie de masse est sans doute l'étape la plus importante de notre approche globale. De la qualité des données acquises dépendra la fiabilité et la justesse de l'information qui pourra en être extraite. Il était donc nécessaire de mettre au point une méthode sur le Synapt G2 qui soit adaptée à l'analyse des protéines intactes.

III.1 Séparation UPLC

III.1.1 Généralités

Nous avons rappelé lors d'un précédent chapitre les nombreux avantages qu'offre l'UPLC par rapport à l'HPLC classique. La chaîne chromatographique Acquity UPLC du constructeur Waters dont nous disposons au laboratoire offre en premier lieu un gain significatif de résolution grâce au faible diamètre des particules (< 2 µm) constituant les colonnes UPLC compatibles. La possibilité de travailler à très haute pression (200-1200 bars) permet de raccourcir significativement la durée des analyses et donc d'atteindre un débit d'analyse plus important. Par exemple, lorsque l'analyse d'un mélange de protéines intactes en HPLC nécessitera plus de 60 minutes, la même analyse pourra être effectuée en moins de 20 minutes en UPLC[228].

La taille et l'hydrophobicité relative des histones donnent lieu à une multitude de site d'adsorption sur la colonne qui pourrait faire craindre qu'elles n'y restent accrochées pour finir par la colmater. Le fait de travailler à des pressions élevées en UPLC permet d'éviter que les protéines entières ne restent fixées sur la colonne entre deux analyses[229] (phénomène de *carryover*) et améliore la récupération des espèces les plus hydrophobes. La réduction du temps d'analyse implique également que les protéines sont en contact moins longtemps avec la colonne, réduisant ainsi les phénomènes d'hydrolyse acide sur colonne[230].

L'UPLC se prête donc très bien à l'analyse de protéines intactes. Une méthode de séparation des histones sur une chaîne Accela UHPLC, l'équivalent chez le constructeur Thermo Scientific de l'Acquity UPLC, a d'ailleurs été proposée pour la

première fois par Contrepois *et al.*[176] en 2010. Les auteurs sont ainsi parvenus à réaliser une séparation chromatographique d'histones en mélange extraites de fibroblastes WI-38 en seulement 19 minutes. En se servant de cette méthode comme point de départ, nous l'avons adaptée à notre système chromatographique et réalisé quelques améliorations présentées dans ce chapitre. Nous présenterons ici les différents choix qui ont conduit à la mise au point de notre méthode chromatographique. Le résumé des paramètres chromatographiques retenus est présenté en partie expérimentale.

III.1.2 Mise au point de la méthode chromatographique

La première étape fondamentale lors du développement d'une méthode chromatographique est le choix de la colonne. Pour l'analyse des histones intactes, nous avons choisi la colonne Waters à polarité de phase inversée Acquity UPLC BEH C_{18} de dimensions 2,1 x 150 mm, contenant des particules de diamètre 1,7 µm et de porosité 300 Å. Nous avons choisi la colonne la plus longue afin d'obtenir la meilleure résolution possible. La mise au point des paramètres chromatographiques a été faite dans le but d'obtenir une méthode offrant le meilleur compromis entre résolution, sensibilité et rapidité de l'analyse.

Phases mobiles :

Deux phases mobiles différentes ont été utilisées. La première est constituée d'eau. La seconde est constituée d'un solvant organique : l'acétonitrile (ACN). C'est un solvant aprotique polaire très utilisé en chromatographie pour la séparation des peptides et des protéines. Sa viscosité relativement faible par rapport à d'autres solvants organiques tels que l'isopropanol permet de réduire la pression en tête de colonne et de prolonger la durée de vie de la colonne, conformément aux recommandations du constructeur qui préconisent de ne pas dépasser 700 bar. En chromatographie liquide à polarité de phase inversée, l'ajout d'un acide dans les phases mobiles est très courant. Cet acide est généralement utilisé pour acidifier les protéines et favoriser la protonation des fonctions carboxylates, et sert également d'agent de paire d'ions pour les résidus basiques afin d'améliorer leur interaction avec la phase stationnaire de la colonne[231]. L'agent le plus couramment utilisé reste l'acide trifluoroacétique (TFA) dont

l'ajout offre une amélioration de la résolution chromatographique. Cependant, il est néfaste pour le spectromètre de masse et est responsable de la formation d'adduits avec les protéines les plus basiques. Dans le cas d'un couplage avec un spectromètre de masse, ces adduits entrainent une suppression de l'ionisation et limitent la détection des espèces les moins abondantes, conduisant à une perte de sensibilité notoire. De plus, ils compliquent davantage l'interprétation des spectres, particulièrement dans le cas des histones où de nombreuses espèces différentiellement modifiées peuvent être co-éluées. Nous l'avons donc remplacé par l'acide formique malgré une diminution de la résolution chromatographique due à la perte de l'effet paire d'ions.

Débit et gradient :

Le débit de phase mobile est un paramètre important qui influence directement la résolution et l'efficacité de la séparation chromatographique. Il doit être choisi en fonction du système utilisé et des dimensions de la colonne. Comme décrit dans la méthode de Contrepois *et al.,* le débit de départ a été fixé à 0,3 mL/min. Un débit plus lent conduirait à une baisse de la pression en tête de colonne ainsi qu'à un élargissement des pics chromatographiques. Nous avons donc testé des débits plus rapides afin de vérifier s'ils offraient une meilleure efficacité de séparation. Il faut toutefois veiller à ne pas avoir un débit trop élevé afin de ne pas introduire une trop grande quantité d'ACN dans la source d'ionisation, ce qui accélèrerait la désolvatation et réduirait l'efficacité d'ionisation[232]. Trois débits ont ainsi été comparés : 0,3 mL/min (figure 55A), 0,4 mL/min (figure 55B) et 0,5 mL/min (figure 55C).

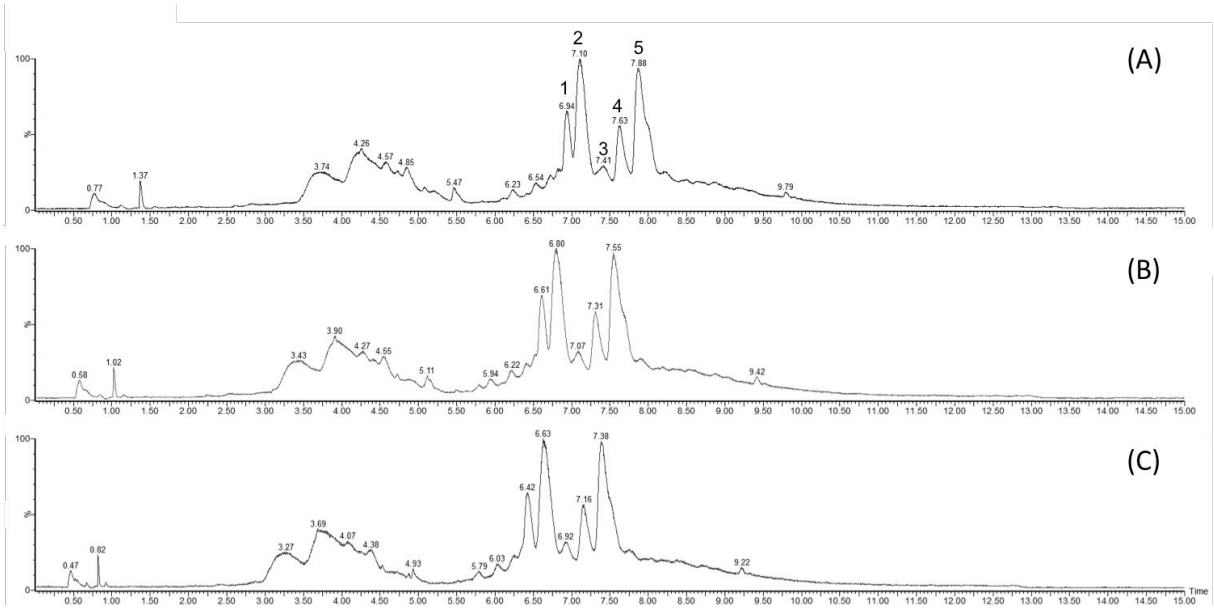


Figure 55 : influence du débit de phase mobile sur l'efficacité de séparation chromatographique d'un mélange d'histones commerciales extraites de thymus de veau. (A) débit = 0,3 mL/min, (B) = 0,4 mL/min, (C) = 0,5 mL/min.

D'après la figure 55, l'augmentation du débit de phase mobile ne semble pas améliorer l'efficacité de la séparation chromatographique : le nombre et la largeur des pics restent identiques. A l'inverse, plus le débit est élevé plus l'intensité du signal est faible. Le débit à 0,3 mL/min offre la meilleure résolution et la meilleure sensibilité. Il présente également comme avantage de réduire la consommation de solvants.

Ensuite, différentes pentes du gradient en ACN ont été testées. Dans tous les cas, un gradient linéaire a été utilisé (courbe 6, figure 56).

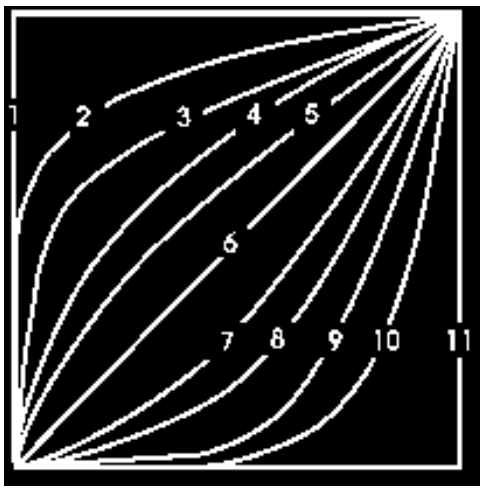


Figure 56 : courbes de gradient proposées sur le système Acquity UPLC du constructeur Waters. Courbe 1 : atteints immédiatement les conditions spécifiées. Courbes 2 à 5 : convexes. Courbe 6 : linéaire. Courbes 7 à 10 : concaves. Courbe 11 : maintien les conditions initiales jusqu'à l'étape suivante.

Le pourcentage initial de phase mobile B (100% ACN) a été fixé à 15% au regard de l'hydrophobicité des histones. Le pourcentage final de B a été fixé à 40%, au-delà duquel tous les sous-types d'histones ont été élués d'après les chromatogrammes. Le paramètre que nous avons fait varier a été la pente du gradient d'ACN. Le gradient a été découpé en trois phases : une phase de 15% à 40% de B ou phase d'élution, une phase de 40% à 60% de B ou phase de rinçage, puis une phase de rééquilibrage de la colonne à 15% de B. Pour la première phase de 15% à 40% de B, trois pentes de gradient ont été comparées : 2,0% B / min (figure 57A), 1,8% B / min (figure 57B) et 1,5% B / min (figure 57C).

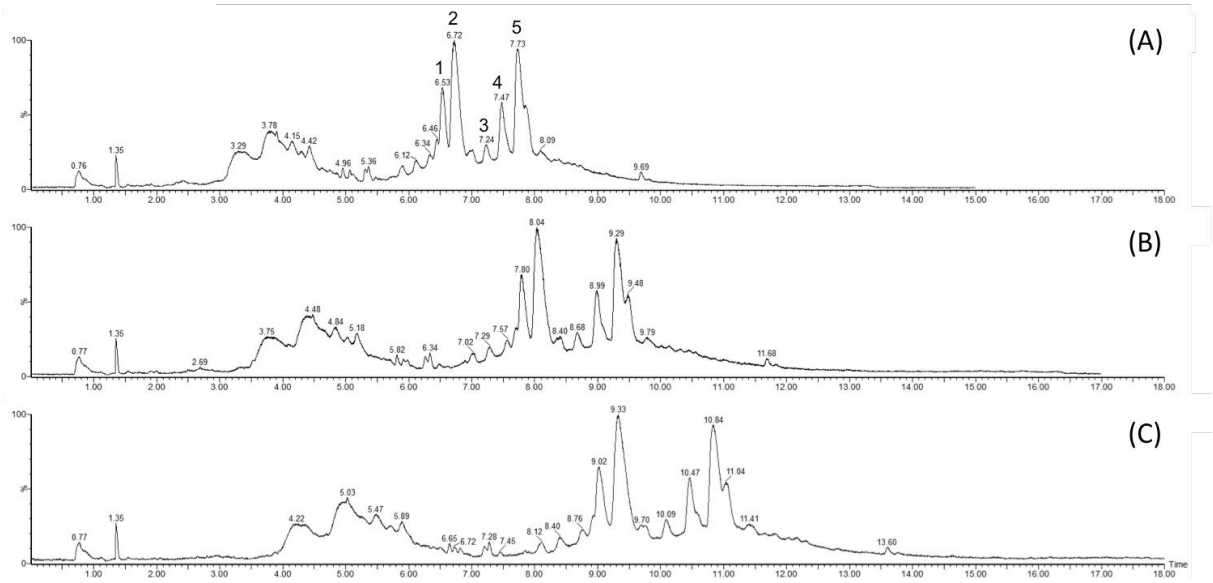


Figure 57 : influence de la pente du gradient linéaire de 15% à 40% de B sur la séparation des histones en mélange. (A) pente = 2,0% B / min, (B) pente = 1,8% B / min, (C) pente = 1,5% B / min.

Généralement, une pente de gradient plus faible offre une meilleure résolution au détriment de la sensibilité. La résolution $R_{1/2}$ a donc été évaluée entre le pic 1 (histone H4) et le pic 2 (histone H2B) pour chaque pente de gradient (tableau 10).

Tableau 10 : résolution de la séparation chromatographique entre les pics 1 et 2 en fonction de la pente du gradient.

Pente (% B / min)	2,0	1,8	1,5
Résolution $R_{1/2}$	1,23	1,41	1,68

La meilleure résolution a donc été obtenue avec la pente à 1,5 % de B / min. La perte de sensibilité associée à cette pente plus faible est moindre par rapport au gain de résolution et n'empêche pas de détecter le même nombre d'espèces.

Température de colonne :

La chaîne chromatographique Acquity dispose d'un four à colonne thermostaté. Une température de colonne élevée offre généralement une meilleure récupération des protéines et parfois une meilleure séparation chromatographique[228]. Cependant une température trop élevée réduit la durée de vie de la colonne et risquerait de dégrader les protéines. Il est généralement recommandé par le constructeur de ne

pas dépasser une température de 65°C sur les colonnes Acquity UPLC. La température du four initiale a été fixée à 50°C comme décrit par Contrepois *et al.* puis elle a été augmentée par paliers de 5°C jusqu'à atteindre 65°C (figure 58).

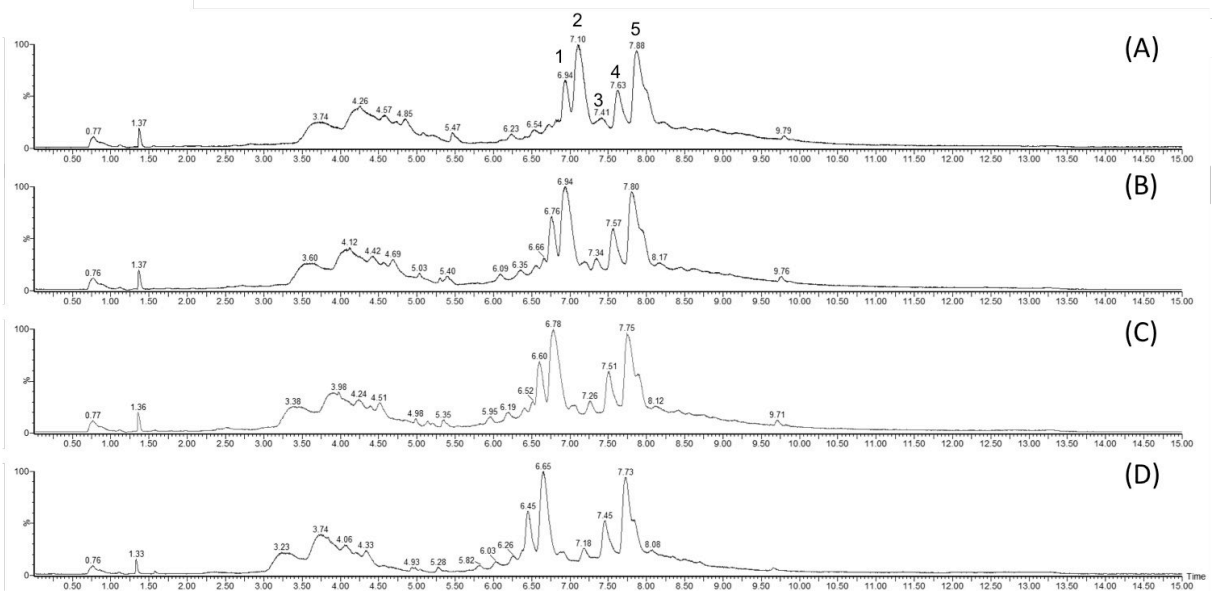


Figure 58 : effet d'une augmentation de la température (T) de colonne sur la séparation chromatographique des histones en mélange. (A) T = 50°C, (B) T = 55°C, (C) T = 60°C, (D) T= 65°C.

D'après la figure 58, l'augmentation de température de colonne semble légèrement améliorer la résolution chromatographique, mais au détriment d'une perte de sensibilité significative. Les pics sont également un peu plus fins avec une température plus élevée. Nous avons estimé que l'amélioration observée ne justifiait pas de dépasser une température de 55°C.

Linéarité :

Pour évaluer la linéarité de la méthode chromatographique, nous avons injecté différentes quantités d'histone H4 humain recombinante puis mesuré l'aire sous le pic chromatographique. La corrélation entre la quantité de protéines injectée et l'aire sous le pic doit être la meilleure possible. L'objectif étant d'appliquer cette méthode a des échantillons précieux, l'analyse doit pouvoir se faire en utilisant la plus petite quantité de protéines possible, tout en conservant une sensibilité suffisante pour détecter les espèces peu abondantes. La linéarité a donc été testée pour des quantités de protéine allant de 0,5 à 2,0 µg. En dessous de 0,5 µg injecté certains pics correspondant à des espèces peu abondantes ne sont pas détectés. Au

delà de 2,0 µg injectés, le gain d'information ne semble pas justifier la surconsommation de l'échantillon. Pour chaque quantité injectée, un triplicat analytique a été réalisé (tableau 11 et figure 59).

Tableau 11 : linéarité de la méthode chromatographique évaluée en injectant des quantités croissantes d'histone H4 humaine recombinante (en triplicat). La moyenne, l'écart-type et le coefficient de variation des aires sous le pic sont présentés.

Quantité (µg)	0,5	0,75	1	1,5	2
Moyenne	57757,67	80849,67	101524,67	133707,67	161010,33
Écart-type	744,10	1288,21	1463,56	962,68	988,30
CV (%)	1,29	1,59	1,44	0,72	0,61

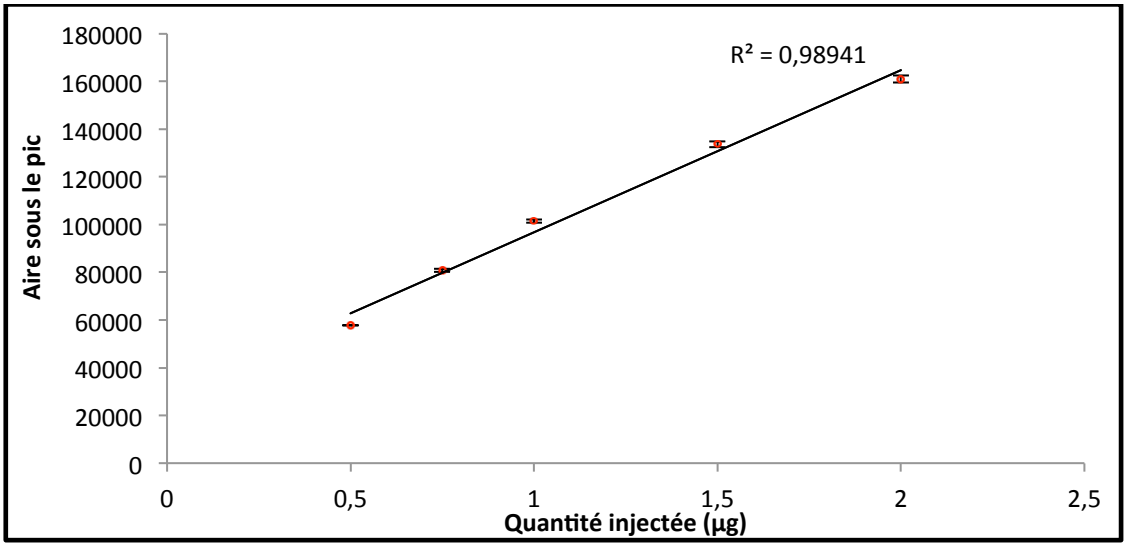


Figure 59 : droite de régression linéaire évaluant la linéarité de la méthode chromatographique.

La droite de régression linéaire nous fournit un coefficient de détermination (R^2) de 0,989, ce qui est acceptable et nous permet d'affirmer que la méthode est linéaire. Autrement dit, l'aire sous le pic chromatographique est bien proportionnelle à la quantité de protéines injectée dans la gamme de quantités testées.

Répétabilité :

Toujours dans l'objectif de prouver la validité de la méthode, nous avons cherché à évaluer sa répétabilité. Ce paramètre est particulièrement important dans la mesure où nous travaillons sur des protéines entières qui s'adsorbent davantage sur la colonne et dont le comportement chromatographique ne sera pas aussi constant que celui de peptides. Ainsi, la stabilité des temps de rétention (tableau 12) et des aires sous les pics chromatographiques (tableau 13) a été évaluée en injectant successivement cinq réplicats analytiques. Afin de s'assurer que la méthode est répétable quel que soit le type d'histone de coeur étudié, nous avons choisi d'évaluer la répétabilité sur chacun de ces sous-types à partir du mélange d'histones commerciales de veau. La stabilité des temps de rétention des 4 sous-types d'histones de cœur est excellente. D'après le tableau 12, les coefficients de variation (CV) sont inférieurs à 0,6% dans tous les cas. Ceci signifie que la stabilité des temps de rétention est excellente et qu'il n'y a pas de dérive au cours du temps.

Tableau 12 : répétabilité des temps de rétention évaluée sur chacun des sous-types d'histones de cœur à partir d'un mélange d'histones commerciales.

	Temps de rétention			
N° réplicat	**H4**	**H2B**	**H2A**	**H3**
1	8,81	9,31	10,97	11,64
2	8,82	9,31	10,97	11,63
3	8,81	9,31	10,97	11,64
4	8,73	9,26	10,9	11,56
5	8,71	9,23	10,89	11,56
Moyenne	8,78	9,28	10,94	11,61
Ecart-type	0,05	0,04	0,04	0,04
CV (%)	0,59	0,40	0,38	0,36

La stabilité des aires sous les pics chromatographiques a été évaluée à partir des mêmes chromatogrammes, et les résultats sont résumés dans le tableau 13. Les coefficients de variation sont ici inférieurs à 4,0% dans tous les cas, et les aires sous les pics sont également très stables au cours du temps.

Tableau 13 : répétabilité des aires sous les pics chromatographiques évaluée sur chacun des sous-types d'histones de cœur à partir d'un mélange d'histones commerciales.

	Aire sous le pic			
N° réplicat	**H4**	**H2B**	**H2A**	**H3**
1	52698	313842	49704	345475
2	54705	313837	49286	347884
3	55204	315183	45649	348988
4	55367	313639	48641	345456
5	53933	307542	46164	343357
Moyenne	54381	312809	47889	346232
Ecart-type	1094	3008	1858	2223
CV (%)	2,0	1,0	3,9	0,7

La répétabilité des temps de rétention et des aires sous les pics chromatographiques atteste de la robustesse du système UPLC utilisé. Nous avons également choisi le mode d'injection *partial loop* qui ne prélève que le volume directement injecté afin d'éviter de consommer un excès d'échantillon pour rincer la boucle d'injection. Au final, la méthode mise au point sur des histones commerciales extraites de thymus de veau a été testée sur un mélange d'histones extraites de cellules BeWo (figure 60). Avec les conditions chromatographiques établies, nous sommes parvenus à obtenir une séparation des différentes histones très satisfaisante au regard de leurs caractéristiques physico-chimiques très proches et de la complexité du mélange.

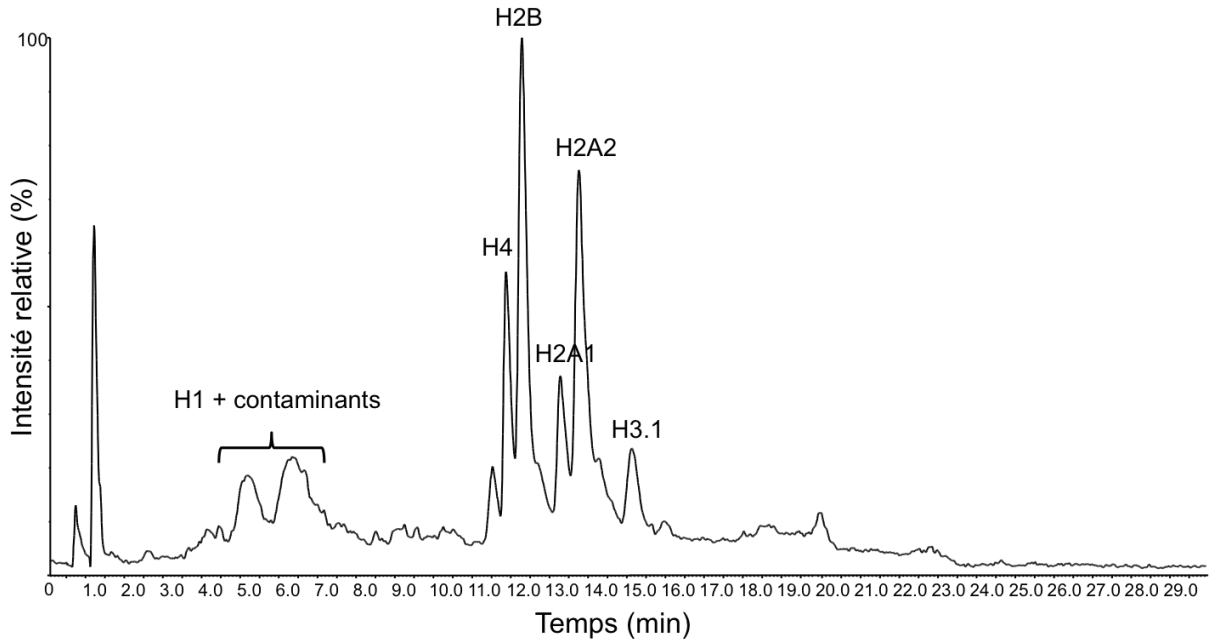


Figure 60 : chromatogramme de référence obtenu à partir de l'injection de 1,5 µg d'histones extraites de cellules BeWo.

III.2 Spectrométrie de masse ESI-QTOF

Une fois les paramètres de chromatographie choisis, il faut mettre au point les paramètres propres au spectromètre de masse. Dans le cadre d'une analyse de protéines intactes, les paramètres les plus importants concernent la source d'ionisation. En effet, du bon réglage des paramètres de la source dépendra l'efficacité d'ionisation, la détection des espèces présentes et *in fine* la qualité des données acquises. Les paramètres du détecteur utilisés sont les paramètres standards indiqués par le constructeur. Seules les mises au point des paramètres de source et de l'analyseur TOF ont été effectuées et seront présentées au cours de ce chapitre.

III.2.1 Mise au point des paramètres de source

III.2.1.1 Généralités

La source ESI montée sur le Synapt G2 de Waters possède sept paramètres principaux qu'il est possible de faire varier : tension du capillaire, tension du cône d'échantillonnage, tension du cône d'extraction, température de source,

température de désolvatation, flux d'azote sur le cône d'échantillonnage et flux d'azote pour la désolvatation. Le constructeur indique des valeurs standard pour chacun d'entre eux. Certains de ces paramètres comme les deux flux d'azote dépendent directement du débit et de la composition de la phase mobile. Ces paramètres ont donc été fixés de manière réfléchie d'après les conditions chromatographiques. Nous avons choisi d'évaluer l'impact de quatre paramètres sur la sensibilité de la détection par le spectromètre de masse : la tension du capillaire, celle du cône d'échantillonnage, la température de source et celle de désolvatation.

Ces paramètres ne sont évidemment pas indépendants les uns des autres. Déterminer leur influence sur la sensibilité de détection du spectromètre de masse de façon empirique ne serait pas très pertinent. Nous avons donc choisi d'utiliser un plan d'expérience. Les plans d'expérience permettent de tester un grand nombre de conditions expérimentales de la manière la plus rationnelle possible afin de diminuer au maximum le nombre d'essais.

III.2.1.2 Plan factoriel complet à deux niveaux

Le plan que nous avons choisi de mettre en place est un plan de criblage appelé plan factoriel complet (PFC) à deux niveaux. Ce type de plan d'expérience assez classique permet d'identifier quels sont les facteurs qui ont le plus d'influence sur un paramètre mesuré. Il nécessite de définir un domaine expérimental, c'est-à-dire des bornes pour chacun des paramètres. Il part du principe qu'en cas de réponse linéaire, les points centraux ont moins d'effet de levier que les extrémités. Il permet de tester k facteurs, chacun sur une borne inférieure (notée -) et une borne supérieure (notée +), de façon à générer 2^k essais[233]. En utilisant cette formule, les quatre paramètres que nous souhaitons évaluer nécessitent d'effectuer seize essais différents. Les bornes supérieures et inférieures de chacun de ces paramètres (tableau 14) ont été établies selon les recommandations du constructeur, en veillant à ce qu'elles soient en accord avec l'analyse de protéines intactes.

Tableau 14 : bornes supérieures et inférieures de chacun des facteurs introduits dans le plan factoriel complet à deux niveaux.

Bornes	A tension du capillaire	B tension du cône	C température de source	D température de désolvataion
inférieure (-)	3,0	30	100	300
supérieure (+)	3,5	45	120	600

A partir du domaine expérimental défini, une matrice des essais contenant 16 essais a été générée automatiquement à l'aide du logiciel R (package DoE.base). L'ordre de passage des différents essais a été déterminé aléatoirement afin de limiter le biais dû à l'ordre d'injection appelé effet bloc. Pour chacun des 16 essais, la même quantité d'histone H4 humaine recombinante a été injectée, et l'aire sous le pic du chromatogramme reconstitué (*Extracted Ion Chromatogram*, EIC) de l'ion le plus intense a été utilisée comme paramètre de réponse (tableau 15). Les interactions de premier degré entre les facteurs ont également été prises en compte.

Tableau 15 : matrice des essais du plan factoriel complet à deux niveaux et 4 facteurs. Les réponses pour chaque essai ainsi que les interactions de premier degré entre facteurs sont représentées.

N°	A	B	C	D	Réponse	A*B	A*C	A*D	B*C	B*D	C*D
2	-	-	-	-	61915	+	+	+	+	+	+
14	-	-	-	+	82228	+	+	-	+	-	-
15	-	-	+	-	65013	+	-	+	-	+	-
8	-	-	+	+	84182	+	-	-	-	-	+
5	-	+	-	-	74118	-	+	+	-	-	+
6	-	+	-	+	96003	-	+	-	-	+	-
4	-	+	+	-	73734	-	-	+	+	-	-
16	-	+	+	+	96756	-	-	-	+	+	+
7	+	-	-	-	58608	-	-	-	+	+	+
3	+	-	-	+	77465	-	-	+	+	-	-
10	+	-	+	-	57682	-	+	-	-	+	-
1	+	-	+	+	80667	-	+	+	-	-	+
9	+	+	-	-	70577	+	-	-	-	-	+
13	+	+	-	+	93268	+	-	+	-	+	-
11	+	+	+	-	68658	+	+	-	+	-	-
12	+	+	+	+	93556	+	+	+	+	+	+

A partir de ces résultats, les effets principaux de chaque facteur ont été calculés pour les deux bornes (tableau 16) d'après la formule détaillée en partie expérimentale. Le calcul de ces effets permet de mettre en évidence leur influence sur la réponse, et donc sur la sensibilité de détection.

Tableau 16 : effets principaux de chaque facteur aux bornes supérieures et inférieures. A = tension du capillaire, B = tension du cône d'échantillonnage, C= température de source et D = température de désolvatation.

Facteur	Effet +	Effet -
A	600481	633949
B	666670	567760
C	620248	614182
D	704125	530305

Une représentation graphique simple de ces résultats permet de visualiser l'effet direct d'une variation de chaque paramètre sur la sensibilité (figure 61).

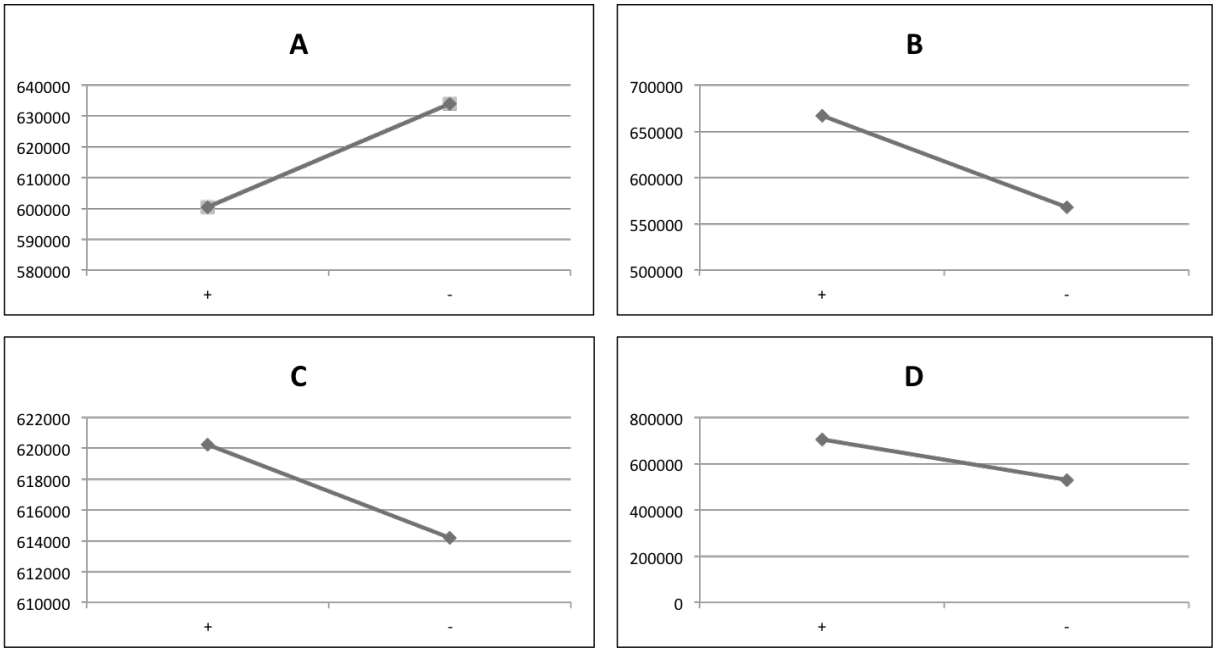


Figure 61 : représentation graphique des effets principaux de chacun des facteurs évalués dans le plan factoriel complet. A = tension du capillaire, B = tension du cône d'échantillonnage, C= température de source et D = température de désolvatation. L'axe des ordonnées représente l'aire sous le pic du chromatogramme reconstitué à partir de l'ion le plus intense.

Les effets moyens ont ensuite été calculés (tableau 17) d'après la formule détaillée en partie expérimentale afin de mieux se rendre compte de l'impact de

chacun des facteurs sur la réponse ainsi que des interactions de premier ordre existant entre eux.

Tableau 17 : effets moyens de chacun des facteurs et de leurs interactions de premier ordre.

Facteur	Effet moyen
A	-4183,5
B	12363,75
C	758,25
D	21727,5
A*B	545,5
A*C	-597
A*D	630,25
B*C	-1073,75
B*D	1396,5
C*D	791

Pour évaluer l'erreur expérimentale sur la mesure de la réponse, il est impératif de disposer d'un point central sur lequel la mesure sera répétée cinq fois. Les valeurs utilisées pour le point central sont les valeurs moyennes de chaque facteur se situant au milieu des deux bornes (tableau 18).

Tableau 18 : valeurs de chaque facteur du point central.

	A	**B**	**C**	**D**
Facteur	tension du capillaire	tension du cône	température de source	température de désolvataion
Point central	3,25	38	110	450

En injectant le même échantillon cinq fois (tableau 19), il est possible de calculer une erreur expérimentale relative (CV) sur la mesure qui sera extrapolée à toutes les mesures faites précédemment.

Tableau 19 : évaluation de l'erreur relative sur la mesure de la réponse à partir de cinq réplicats analytiques du point central.

	Réponse
R1	82394
R2	79653
R3	80660
R4	79469
R5	81672
Moyenne	80769,6
Ecart-type	1265,2
CV (%)	1,57

Le coefficient de variation (CV) calculé à partir des réplicats du point central est appliqué aux effets moyens calculés précédemment. Le diagramme en bâton présenté à la figure 62 permet de visualiser ces effets moyens avec l'erreur expérimentale.

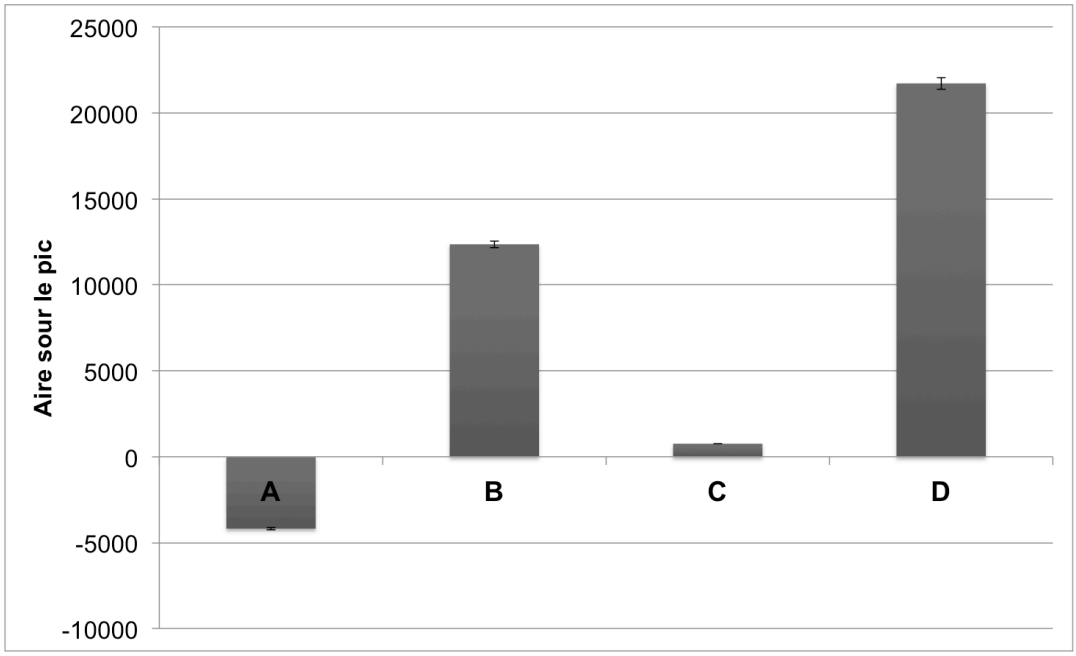


Figure 62 : effets moyens des quatre facteurs sur la sensibilité du spectromètre de masse. Les barres d'erreurs représentent l'erreur relative calculée à partir des réplicats du point central. A = tension du capillaire, B = tension du cône d'échantillonnage, C= température de source et D = température de désolvatation.

Ce type de représentation permet de déduire intuitivement l'effet des différents facteurs sur la sensibilité de l'appareil. Nous pouvons ainsi affirmer

qu'une augmentation de la tension du cône d'échantillonnage, de la température de source et de la température de désolvatation ont, à une échelle différente, un effet positif sur la sensibilité de l'appareil. A l'inverse, une augmentation de la tension du capillaire induit une perte de sensibilité.

Les résultats observés graphiquement ont été confirmés en réalisant une régression linéaire multiple à l'aide du logiciel R. Cette méthode permet d'évaluer la significativité des effets et des interactions en réalisant un test de Student sur chaque facteur (figure 63) afin de déterminer s'il est linéairement associé à la réponse. D'après ces résultats, tous les facteurs, excepté la température de source ont un effet significatif sur la sensibilité. En ce qui concerne les interactions, seule l'interaction entre la tension du cône d'échantillonnage et la température de désolvatation est légèrement significative.

```
lm.default(formula = Response ~ (Capillary + Sampling + SourceT +
    Desolvatation)^2, data = Design.1.withresp)

Residuals:
      1       2       3       4       5       6       7       8       9
10
  82.88 -612.87  653.13 -184.63  224.88 -664.87 -750.13 -127.88  -42.63 -
179.88
     11      12      13      14      15      16
-608.13  139.63  705.13  709.87   87.63  567.87

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               77151.9      214.2 360.233 3.13e-12 ***
Capillary1                -2091.8      214.2  -9.767 0.000191 ***
Sampling1                  6181.9      214.2  28.864 9.35e-07 ***
SourceT1                    379.1      214.2   1.770 0.136914
Desolvatation1            10863.8      214.2  50.724 5.63e-08 ***
Capillary1:Sampling1        272.7      214.2   1.274 0.258828
Capillary1:SourceT1        -298.5      214.2  -1.394 0.222173
Capillary1:Desolvatation1   315.1      214.2   1.471 0.201163
Sampling1:SourceT1         -536.9      214.2  -2.507 0.054043 .
Sampling1:Desolvatation1    698.2      214.2   3.260 0.022440 *
SourceT1:Desolvatation1     395.5      214.2   1.847 0.124083
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 856.7 on 5 degrees of freedom
Multiple R-squared:  0.9986,  Adjusted R-squared:  0.9958
F-statistic: 353.1 on 10 and 5 DF,  p-value: 1.757e-06
```

Figure 63 : résultats de la régression linéaire multiple effectuée à l'aide du logiciel R. Ces résultats montrent les valeurs résiduelles de chaque essai par rapport à la droite de régression. Le test de Student indique à l'aide d'astérisques la significativité des effets et des interactions.

La faible valeur de *p-value* fournie par l'analyse de variance (ANOVA) ainsi que la valeur élevée du test de Fischer prouvent que la variance de régression et la variance résiduelle sont significativement différentes, attestant ainsi de la qualité de la régression. Le plan factoriel complet réalisé nous a donc permis d'identifier les paramètres influençant significativement la sensibilité de détection du spectromètre de masse. Ces paramètres ayant une réponse linéaire, nous n'avons pas jugé utile de réaliser un plan d'optimisation par la suite. Pour établir la méthode finale, nous avons donc choisi les paramètres pour lesquels la réponse était maximale. Ces paramètres sont résumés en partie expérimentale.

III.2.2 Réglage de la fréquence d'acquisition de l'analyseur TOF

En plus des paramètres de source, la fréquence d'acquisition de l'analyseur TOF est un paramètre fondamental qu'il faut ajuster selon les conditions chromatographiques pour obtenir des données de qualité. La fréquence d'acquisition correspond au nombre de mesure de rapports *m/z* par seconde effectuées par l'analyseur. Elle dépend directement des conditions chromatographiques, et plus particulièrement de la largeur des pics. Si elle est mal réglée, cela peut entraîner une augmentation du bruit de fond et une perte significative de sensibilité. Plus un pic est large moins l'analyseur devra avoir une fréquence d'acquisition élevée. En effet, le nombre minimum de points nécessaires pour définir un pic gaussien est estimé à environ 20. Au-delà, la quantité d'information acquise augmente au détriment de la sensibilité. Il suffit donc d'évaluer la largeur du pic chromatographique le plus étroit pour en déduire la fréquence d'acquisition optimale. Dans le cas d'un mélange d'histones extraites de cellules BeWo, la largeur du pic le plus fin est de 10 secondes. S'il nous faut 20 points pour définir ce pic, l'analyseur TOF devra acquérir 2 points par seconde. La fréquence d'acquisition devra donc être de 2 Hz.

Sur le spectromètre de masse Synapt G2 HDMS, cette fréquence est relativement faible. Généralement, la largeur des pics chromatographiques en UPLC est bien moindre, ce qui nécessite une fréquence d'acquisition élevée, entre 5 et 10 Hz. Nous avons donc évalué l'effet de la fréquence d'acquisition sur la qualité des données acquises et sur la sensibilité de l'appareil. Pour cela, nous avons comparé les EIC de l'ion le plus intense sur les spectres de l'histone H4 d'un

mélange extrait de cellules BeWo à trois fréquences d'acquisition différentes : 10 Hz, 5 Hz et 2 Hz (figure 64).

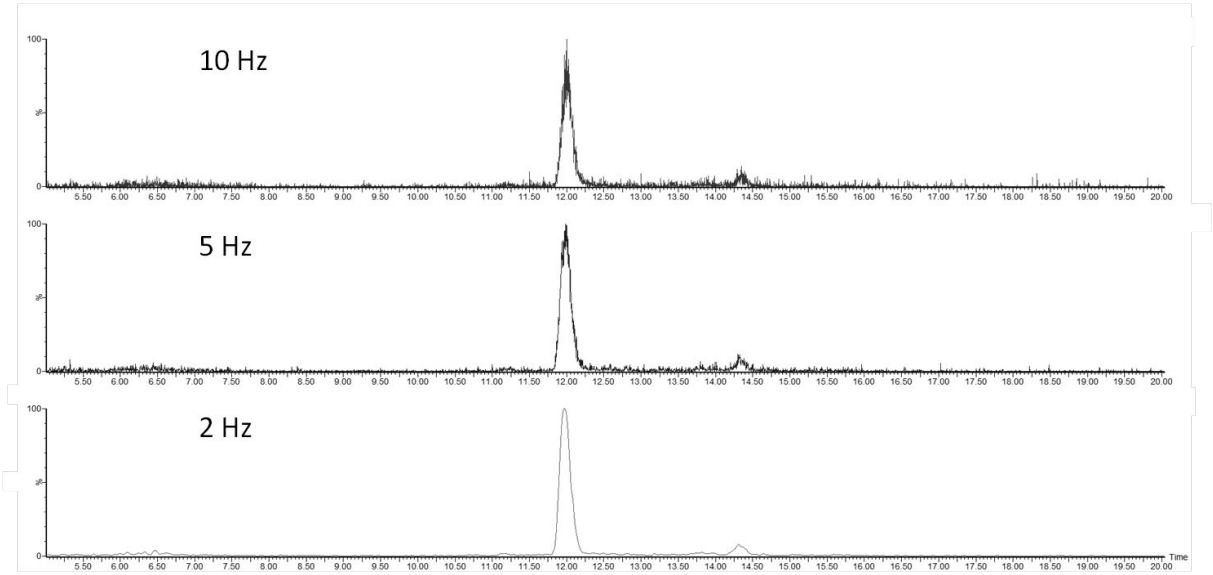


Figure 64 : comparaison de chromatogrammes reconstitués de l'ion m/z 757,566 à une fréquence d'acquisition de 10 Hz, 5 Hz et 2 Hz.

D'après la figure 64, l'augmentation de la fréquence d'acquisition entraîne en parallèle une augmentation nette du bruit de fond. Les fréquences de 5 Hz et 10 Hz semblent bien trop élevées et non adaptées aux conditions chromatographiques. De plus, en intégrant les aires sous les pics, il s'avère que l'augmentation de la fréquence d'acquisition s'accompagne d'une perte de sensibilité quasi linéaire. La meilleure fréquence d'acquisition dans notre cas est donc de 2 Hz.

III.2.3 Continuum *versus* centroïde

Sur le Synapt G2 HDMS, deux modes d'acquisition des données sont disponibles : le mode continuum et le mode centroïde. Le mode continuum représente un signal distribué sur l'ensemble des valeurs *m/z* détectées pour chaque ion. Le mode centroïde, lui, ne retient lui que le maximum local de la distribution des valeurs *m/z* détectées pour un ion (figure 65). Autrement dit, le mode centroïde correspond à une simplification des données mais peut en contre-partie aboutir à une perte d'information.

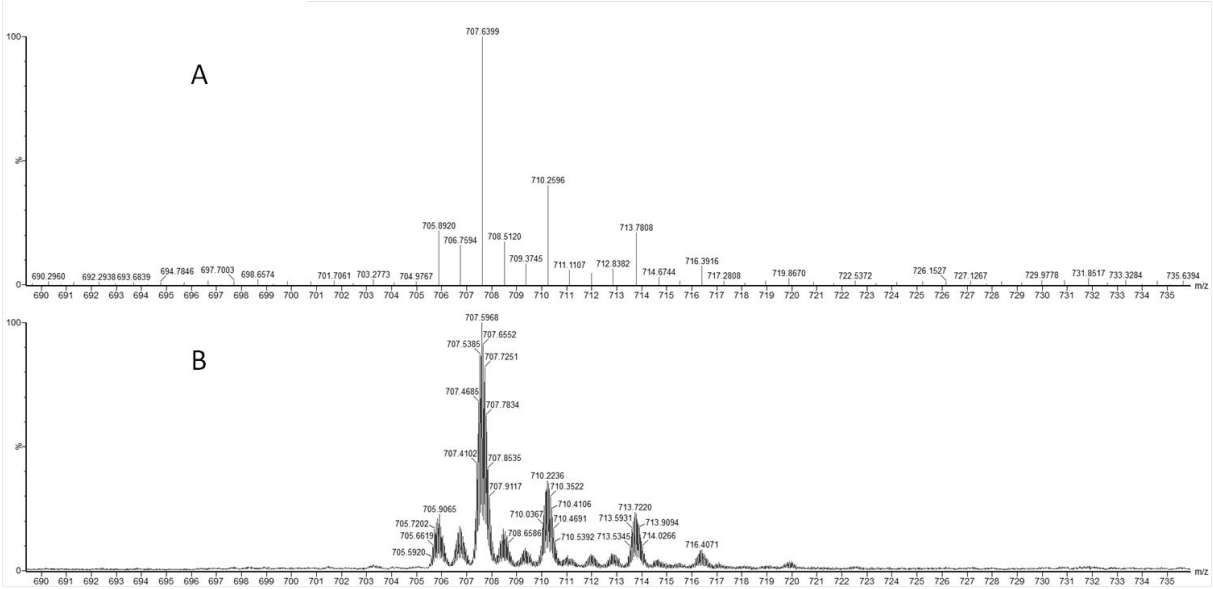


Figure 65 : comparaison de deux spectres d'un même échantillon d'histone H4 humaine recombinante acquis en mode centroïde (A) et en mode continuum (B).

Dans le cas de l'analyse de protéines intactes en ESI, nous avons vu que les spectres présentaient des ions multichargés. Le calcul automatique de la masse moyenne d'une protéine à partir de son spectre ESI ne peut se faire que si les données ont été acquises en mode continuum. D'autre part, le prétraitement des données en vue d'une analyse statistique présentée ultérieurement nécessite de disposer de spectres acquis en mode centroïde. C'est pourquoi nous avons décidé d'acquérir nos données avec les deux modes successivement. Les spectres acquis en mode centroïde seront utilisés pour comparer les profils d'histones et les données en mode continuum seront utilisées pour calculer les masses moyennes des protéines et interpréter les résultats.

Nous avons donc détaillé dans ce chapitre la mise au point de la stratégie de profilage des histones intactes par LC-MS. Les données acquises à l'aide de cette stratégie ne pourront pas être traitées de la même façon que les données de protéomique classique puisque nous avons pris le parti de ne pas réaliser de fragmentation MS/MS mais d'exploiter au maximum les données MS. L'objectif final de ce travail étant de comparer les profils d'histones de cellules placentaires exposées ou non à des polluants environnementaux, la stratégie d'analyse des données devra intégrer simultanément l'ensemble des variables détectées afin de révéler d'éventuels marqueurs d'exposition.

III.3 Ordre d'injection et échantillons « contrôle qualité »

Lorsque l'on souhaite comparer des profils d'histones entre plusieurs groupes d'échantillons de façon à en extraire une information statistiquement pertinente, le choix de l'ordre d'injection des différents échantillons est primordial. Afin de garantir une répartition équilibrée des échantillons au cours du temps et éviter les effets de bloc, les échantillons sont équitablement répartis en fonction de leur nature. La liste des échantillons est dite orthogonalisée, c'est-à-dire que la nature des échantillons ne sera pas corrélée à l'ordre d'injection. De cette façon, l'ordre d'injection des échantillons ne représentera pas un facteur discriminant. En pratique, ce principe consiste à construire une série d'échantillons en les alternant selon leur nature, puis à répéter cette série *n* fois en fonction du nombre d'échantillons total. Par ailleurs, un échantillon de contrôle analytique (*quality control*, QC) est injecté à intervalles réguliers tout le long de l'analyse. Il est constitué d'une quantité équivalente de tous les extraits histoniques injectés. C'est donc une sorte de pool représentatif de l'ensemble des échantillons, qui permettra de suivre la stabilité du système et l'intensité des différentes variables tout au long de l'analyse.

IV. Traitement des données LC-MS

La stratégie de profilage des histones intactes par LC-MS produit un très grand volume de données complexes. Plusieurs milliers d'ions sont détectés pour chaque échantillon. L'absence de données MS/MS à ce stade nous prive de l'utilisation des logiciels classiques de protéomique. Il a donc fallu établir une stratégie d'analyse des données permettant de rechercher des marqueurs d'exposition. Avant d'en arriver à l'analyse des données elle-même, les fichiers bruts directement issus du spectromètre de masse doivent être traités pour pouvoir extraire correctement l'information qu'ils contiennent et parvenir à en tirer des conclusions biologiques. Nous détaillerons ainsi dans ce chapitre les étapes qui nous ont permis de transformer les fichiers LC-MS bruts en matrices des variables prêtes à être analysées à l'aide de méthodes statistiques multivariées.

IV.1 Conversion des fichiers

Le format des fichiers bruts dépend du constructeur de l'appareil sur lequel ils ont été générés. Dans le cas d'un appareil Waters, les fichiers sont au format raw. Les outils que nous allons utiliser par la suite pour extraire l'information de ces fichiers et analyser les données nécessitent de disposer de fichiers aux formats libres tels que mzXML, netCDF ou mzData. Le logiciel MassLynx du constructeur Waters utilisé pour ouvrir les fichiers raw dispose d'un module de conversion appelé DataBridge. Il permet de convertir les fichiers raw au format netCDF. Cependant, les spectres acquis sur le Synapt G2 HDMS présentent des interruptions régulières d'acquisition dues à l'infusion d'un composé appelé *lockmass* par la LockSpray™ pour l'étalonnage en continu de l'appareil. Ces interruptions présentes sur le spectre risqueraient de fausser les analyses ultérieures. Le module DataBridge ne prenant pas en compte la LockSpray™, nous avons utilisé une approche développée par Stanstrup *et al.*[234] qui permet de corriger ces interruptions lors de la conversion. Elle consiste à ouvrir les fichiers centroïdes sous l'environnement R afin de détecter les interruptions dues à l'infusion de la *lockmass* et à les remplacer par l'information acquise juste avant. Les fichiers ainsi complétés sont ensuite convertis par le logiciel massWolf (*Seattle Proteome Center*) au format mzData.

IV.2 Prétraitement des données et extraction des signaux

IV.2.1 Généralités

Les données générées en LC-MS se présentent initialement sous la forme d'un chromatogramme pour chaque échantillon. Sur chaque chromatogramme est représenté l'intensité du courant ionique total (*Total Ion Current*, TIC) en fonction du temps de rétention (t_R). Pour chaque temps de rétention, il existe une série d'ions détectés ayant des rapports *m*/*z* différents. Ainsi, chaque ion détecté possède trois caractéristiques : son rapport *m*/*z*, son temps de rétention et l'intensité de son signal (figure 66). La combinaison entre le temps de rétention

d'un ion et sa valeur *m/z*, notée t_R_*m/z* permet de l'identifier et constitue ce que l'on appellera une variable.

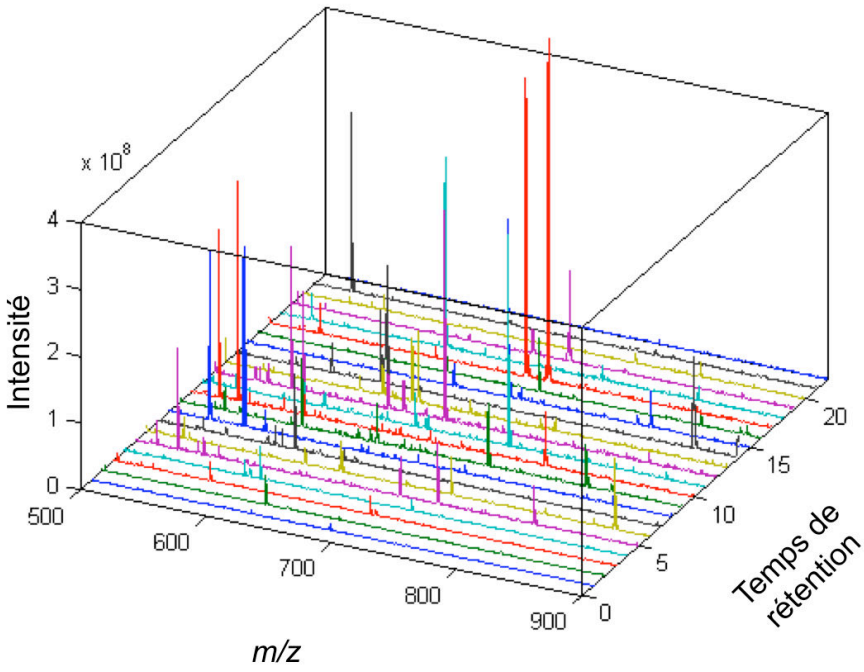
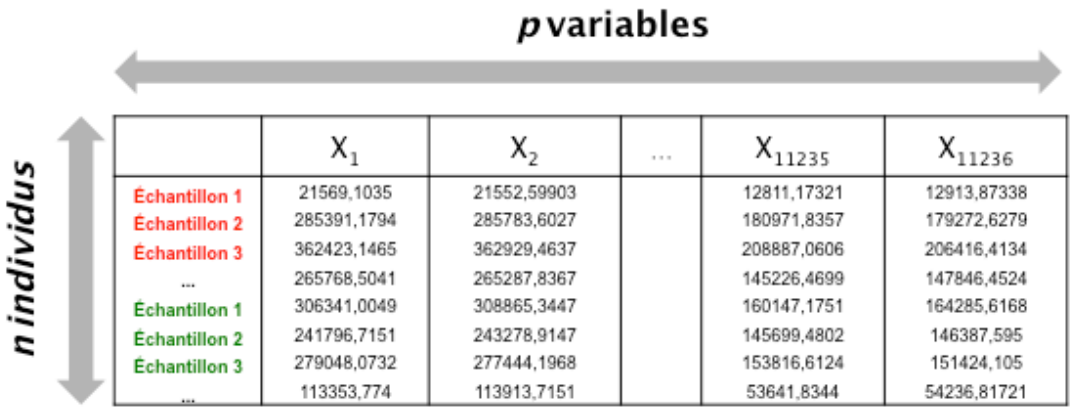


Figure 66 : représentation tridimensionnelle d'un chromatogramme obtenu par LC-MS.

Le prétraitement des données consiste donc à détecter le maximum de variables à partir des chromatogrammes et à extraire les valeurs de leur intensité pour chacun des chromatogrammes afin de pouvoir les comparer. Le logiciel de prétraitement utilisé doit permettre d'éliminer au maximum le bruit de fond pour n'extraire que les signaux correspondant réellement à des ions. Il doit ensuite réaligner et intégrer ces signaux pour créer une matrice notée *X* qui présentera pour *n* individus l'ensemble des variables *p* détectées et leur abondance (figure 67).



	X_1	X_2	...	X_{11235}	X_{11236}
Échantillon 1	21569,1035	21552,59903		12811,17321	12913,87338
Échantillon 2	285391,1794	285783,6027		180971,8357	179272,6279
Échantillon 3	362423,1465	362929,4637		208887,0606	206416,4134
...	265768,5041	265287,8367		145226,4699	147846,4524
Échantillon 1	306341,0049	308865,3447		160147,1751	164285,6168
Échantillon 2	241796,7151	243278,9147		145699,4802	146387,595
Échantillon 3	279048,0732	277444,1968		153816,6124	151424,105
...	113353,774	113913,7151		53641,8344	54236,81721

Figure 67 : exemple d'une matrice de données X représentant les p variables extraites pour n individus appartenant à deux classes différentes (rouge et verte). Chaque variable est identifiée par ses valeurs *m/z* et t_R.

Pour générer de type de matrice, il existe de nombreux logiciels disponibles gratuitement ou commercialement. Nous avons choisi d'utiliser la solution la plus courante actuellement pour le traitement des données LC-MS : la bibliothèque de fonctions (package) XCMS qui fonctionne sous l'environnement R. Ce package présente l'immense avantage d'être accompagné d'une documentation très détaillée ainsi que d'une communauté active d'utilisateurs, ce qui facilite la compréhension des différents algorithmes et le choix des paramètres.

IV.2.2 XCMS

IV.2.2.1 Principe de fonctionnement

La bibliothèque de fonctions XCMS est téléchargeable gratuitement sous R via le site du projet Bioconductor (http://www.bioconductor.org). Le prétraitement sous XCMS fonctionne de manière séquentielle et fait appel à différents algorithmes de détection, d'alignement et d'intégration des signaux. Le principe de fonctionnement de XCMS se décompose en quatre étapes majeures résumées figure 68.

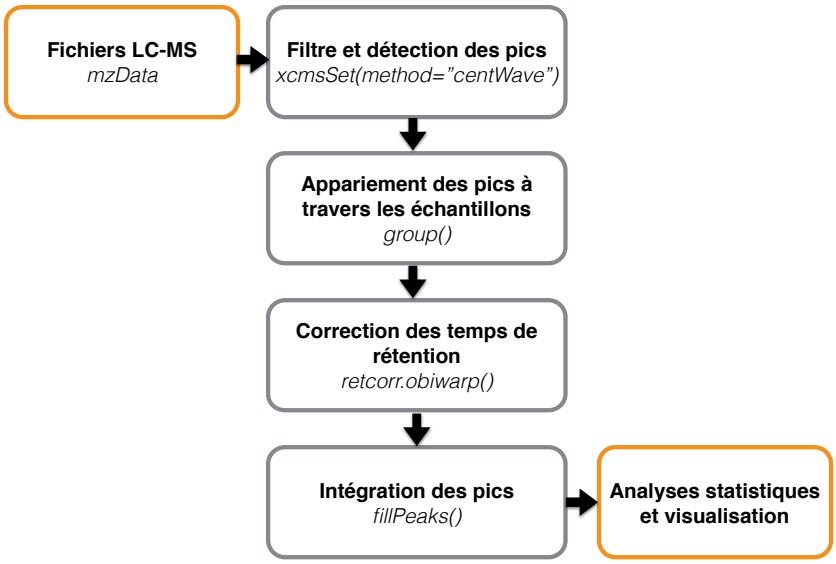


Figure 68 : organigramme des différentes étapes constitutives du prétraitement des données LC-MS par XCMS (en gris). Le package XCMS offre également certains outils statistiques et de visualisation des données.

La première étape détection des pics a été réalisée à l'aide de l'algorithme *centWave* développé par Tautenhahn et ses collaborateurs[235]. Il s'agit d'une évolution de l'algorithme *matchedFilter*[236] implémenté initialement dans le package XCMS. L'algorithme *centWave* a été spécialement conçu pour les données LC-MS à très haute résolution et ne fonctionne qu'avec des données acquises en mode centroïde. Contrairement à l'algorithme *matchedFilter*, l'algorithme *centWave* ne morcèle par les spectres en tranches de rapports *m/z* (*binning*). La première étape consiste à localiser les régions dont la déviation des *m/z* sur des spectres consécutifs est inférieure au paramètre *ppm* (erreur sur la mesure de masse définie par l'utilisateur selon les performances de l'instrument). Ces régions sont appelées régions d'intérêt (figure 69). Une transformée en ondelettes continues (*Continuous Wavelet Transform*, CWT) est ensuite appliquée à travers les régions d'intérêt pour détecter les pics chromatographiques. La détection des EIC tient compte de la largeur moyenne des pics prédéfinie (paramètre *peakwidth)*

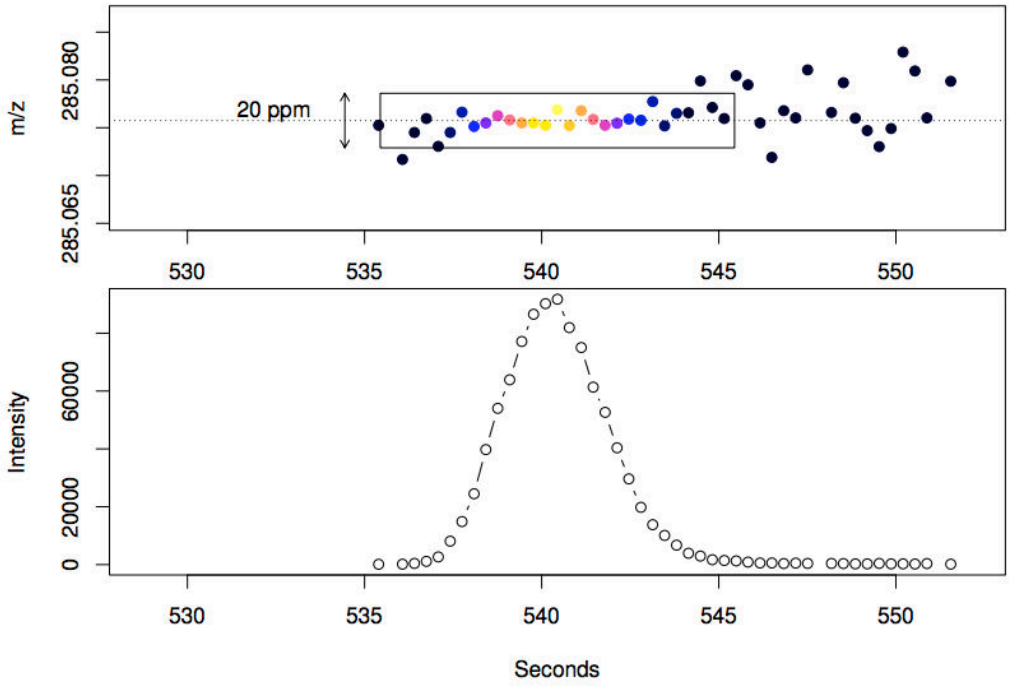


Figure 69 : principe de la détection des régions d'intérêt. Les signaux consécutifs détectés pour l'ion *m/z* 285,075 (haut) dont le rapport *m/z* dévie de moins de 20 ppm constituent une région d'intérêt. Le pic chromatographique correspondant est détecté à l'aide d'une transformée en ondelettes continues. D'après[236].

Après détection de l'ensemble des pics d'intérêt pour chaque échantillon, tous les pics correspondant à une même espèce à travers les échantillons doivent être appariés. Sous XCMS, l'appariement des pics peut se faire à l'aide de différents algorithmes : *group.density*, *group.mzclust* et *group.nearest*. Nous avons choisi d'utiliser la méthode par défaut *group.density*. Cet algorithme tient compte des temps de rétention et des valeurs *m/z* pour apparier les pics. Le paramètre *mzwid* permet de fixer l'intervalle utilisé pour regrouper les pics à l'intérieur d'une tranche de valeurs *m/z (*figure 70).

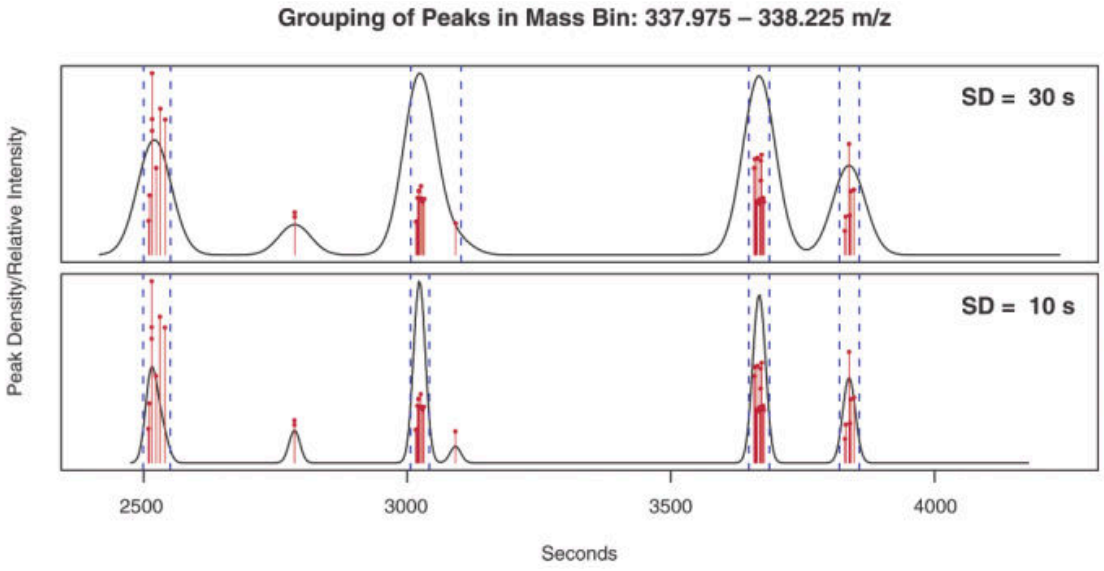


Figure 70 : exemple d'appariement des pics à l'intérieur de la tranche *m/z* 337,975 – 338,225. La valeur de l'écart-type de la gaussienne (haut : SD = 30 s, bas : SD = 10 s) obtenue lors de la phase de lissage impacte la largeur des pics et influence le regroupement à l'intérieur d'une plage de temps de rétention. D'après[236].

Afin d'éviter de découper un pic chromatographique en deux à l'intérieur d'une tranche de valeurs *m/z*, il y a 50% de chevauchement entre deux intervalles successifs ce qui entraîne une redondance de 50% de l'information qui sera supprimée lors de la dernière étape du processus. Une fois les groupes de valeurs *m/z* constitués, ils sont différenciés selon leur temps de rétention. L'intervalle de temps de rétention entre les groupes de *m/z* n'est pas fixe mais est déterminé de manière dynamique par l'algorithme qui identifie les régions contenant de nombreux pics de temps de rétention similaires.

Les groupes de pics étant constitués, l'étape de réalignement les utilise pour identifier et corriger une éventuelle dérive des temps de rétention au fil des injections. Il existe là aussi plusieurs méthodes disponibles sous XCMS : *retcor.loess*, *retcor.linear* et *retcor.obiwarp*. La méthode *obiwarp*[237] a été utilisée dans le cadre de notre travail. Cette méthode calcule la corrélation existant entre les spectres pour estimer leur similitude. S'ils sont similaires, la déviation entre leurs temps de rétention sera corrigée. Les pics ainsi réalignés seront groupés une deuxième fois de manière plus précise.

Enfin, après le deuxième groupement des pics, il peut persister des pics qui n'ont pas été retrouvés dans tous les échantillons. Les intensités de ces pics manquants seront intégrées directement à partir des données initiales.

IV.2.2.2 Utilisation pour l'approche histonomique globale

L'utilisateur dispose d'une très grande flexibilité puisque la totalité des paramètres de XCMS est modifiable. Nous sommes partis de la méthode adaptée aux spectres UPLC-QTOF à haute résolution recommandée par les créateurs du package XCMS puis nous avons modifié quelques paramètres pour qu'elle soit en accord avec les caractéristiques des chromatogrammes (largeurs des pics) et des spectres (résolution, erreur sur la mesure de masse) acquis. Les paramètres choisis sont résumés en partie expérimentale.

Une fois les paramètres déterminés et résumés en partie expérimentale, nous avons voulu tester le prétraitement sur un jeu de données simple contenant deux conditions différentes à comparer. Nous avons ainsi injecté en triplicat deux quantités différentes d'histones commerciales extraites de thymus de veau : 1 et 2 µg. Certaines étapes du processus XCMS ont été évaluées, à savoir l'alignement des chromatogrammes qui permet de s'assurer qu'il s'agit bien du même ion qui est comparé à travers les échantillons, et l'extraction des signaux. Le type de représentation graphique présenté figure 71 permet de vérifier la déviation des temps de rétention.

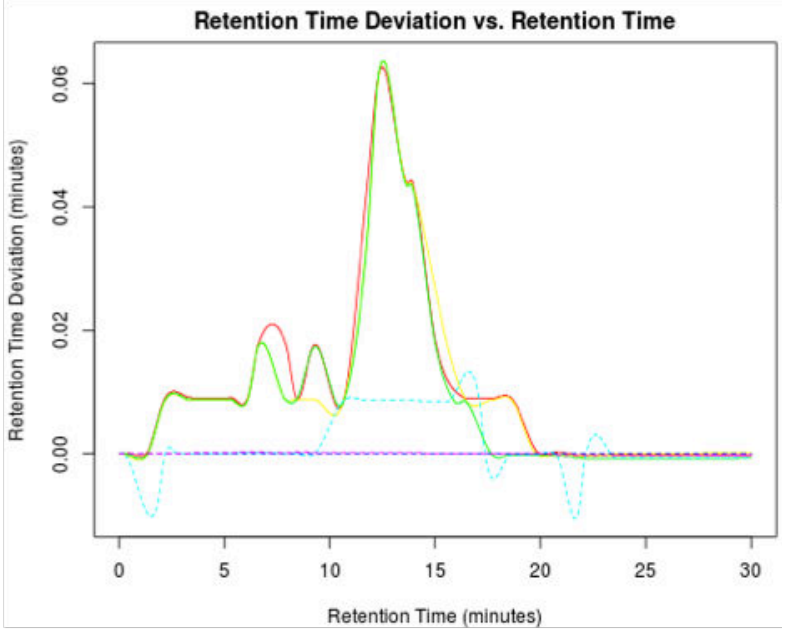


Figure 71 : déviation des temps de rétention (en minutes) pour chacun des échantillons analysés. Chaque trait de couleur représente un échantillon différent.

La très faible valeur des déviations observées confirme la robustesse de notre système chromatographique UPLC. Les chromatogrammes réalignés sont ensuite superposés afin de juger de l'efficacité du prétraitement appliqué (figure 72).

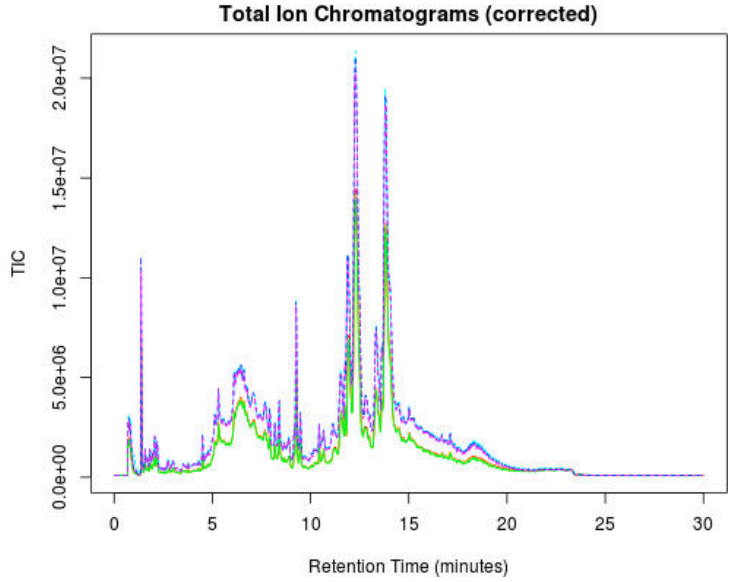


Figure 72 : aperçu de l'ensemble des chromatogrammes (TIC) superposés après réalignement chromatographique réalisé par l'algorithme obiwarp[237].

La figure 73 permet de comparer les chromatogrammes reconstitués (*Extracted-Ion Chromatogram*, EIC) pour un ion défini. Elle nous permet ainsi de constater la qualité de la détection et de l'intégration du signal ainsi que la différence d'aires sous le pic équivalente à un facteur 2 entre les deux groupes d'échantillons.

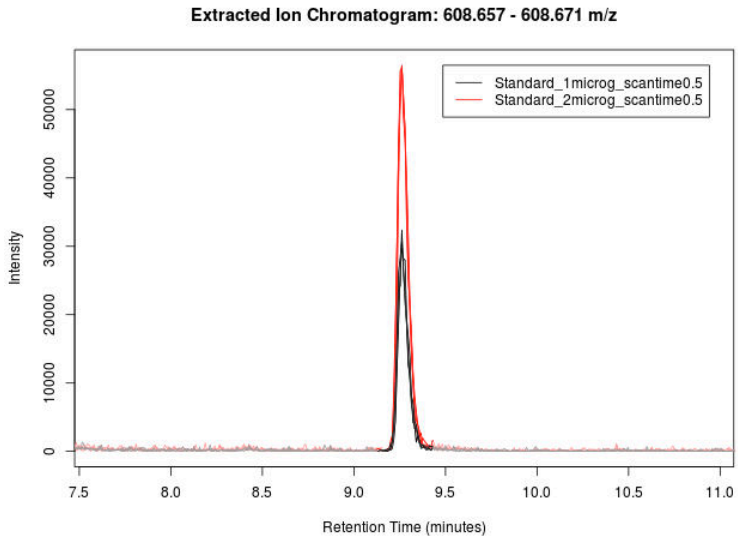


Figure 73 : détection et intégration des signaux correspondants au même ion dans les deux groupes d'échantillons à l'aide de l'algorithme centWave[235].

Une fois le prétraitement par XCMS correctement effectué, une recherche de valeurs manquantes a systématiquement été réalisée sous R sur la matrice *X* des variables. Avant de pouvoir extraire de l'information de cette matrice, elle doit être correctement normalisée afin de limiter au maximum la variabilité indésirable qui risquerait de fausser les résultats des analyses ultérieures.

IV.3 Normalisations

Extraire de l'information biologique pertinente à partir d'un jeu de données complexe représente un véritable défi. En spectrométrie de masse, l'information biologique pertinente est souvent noyée par l'addition d'une variabilité non-induite que l'on appelle du bruit. Cette variabilité non-induite peut être d'origine biologique ou technique (figure 74).

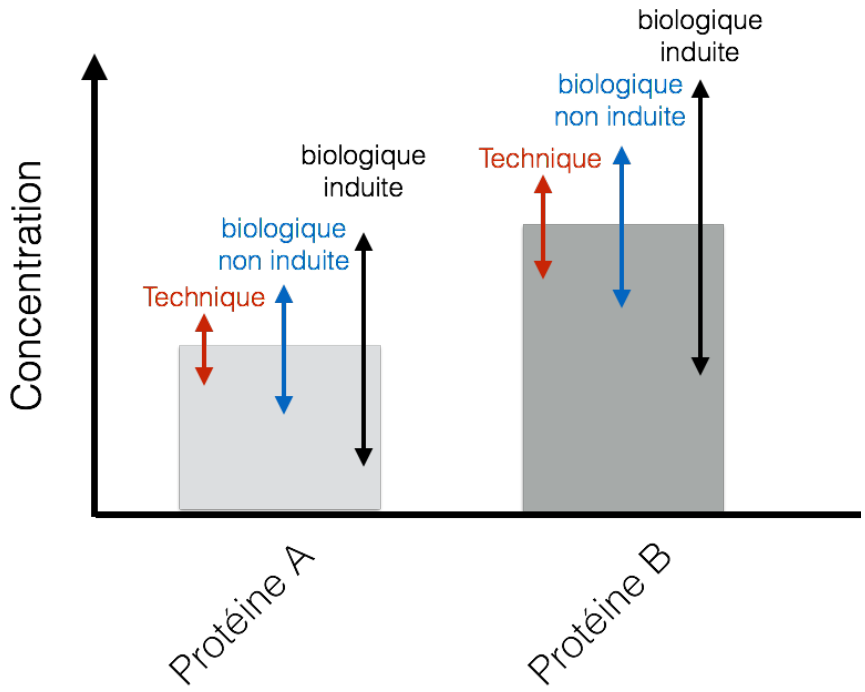


Figure 74 : différents niveaux de variabilité présents dans les données de spectrométrie de masse en biologie. La variabilité totale observée pour une protéine donnée est la somme de la variabilité technique, de la variabilité biologique non induite et de la variabilité biologique induite. La variabilité biologique induite est la seule part de variabilité liée au phénomène biologique étudié. Les autres sources de variabilité sont liées aux conditions expérimentales et sont considérées comme du bruit.

Quelle que soit sa nature ou son origine, cette variabilité indésirable doit être gommée au maximum pour permettre de se focaliser sur la variabilité induite par le phénomène étudié. Pour cela, une normalisation des données à deux niveaux différentes est nécessaire. D'un point de vue statistique, le résultat d'une analyse chimique est considéré comme une variable aléatoire continue. Une démarche probabiliste habituelle doit être employée en vue d'ajuster les résultats expérimentaux à la loi Normale.

Nous présenterons ici les différentes étapes de la stratégie de normalisation des données. Les équations correspondantes sont détaillées en partie expérimentale. Pour illustrer l'effet de chacune de ces étapes, nous les avons appliquées sur le même jeu de données que celui déjà utilisé au chapitre précédent.

IV.3.1 Normalisation inter-échantillons

Dans le cas d'une analyse par spectrométrie de masse, les biais potentiellement introduits au cours de la conception expérimentale ainsi que de l'acquisition des données sont nombreux et difficiles à identifier. Avant de chercher des différences statistiquement significatives entre deux groupes différents d'échantillons biologiques, il faut s'assurer qu'ils sont bien comparables entre eux. En d'autres termes, il faut supprimer au maximum les sources systématiques de variation qui ne sont pas corrélées au phénomène biologique étudié mais aux conditions expérimentales. Ces sources systématiques de variation entre les échantillons sont nombreuses et peuvent être liées à l'échantillon lui-même et/ou à l'instrument. Parmi celles liées à l'échantillon, on rencontre par exemple une variation de la quantité de protéines injectée (dilution) ou une dégradation de certaines protéines au cours du temps. Les sources de variation liées à l'instrument concernent notamment une perte de linéarité de la réponse du détecteur due à une suppression d'ionisation ou un encrassement de la source. Tous ces facteurs introduisent une variabilité indésirable qu'il faut réduire au maximum par une normalisation des données comportant le plus souvent plusieurs étapes. Ces étapes de normalisation doivent cependant préserver la variation biologiquement pertinente que l'on cherche à mettre en évidence.

Il existe classiquement deux stratégies de normalisation inter-échantillons : les approches statistiques et les approches utilisant un ou plusieurs étalons internes[238]. Dans le cas de l'approche histonomique, l'utilisation d'étalons internes étant impossible, nous n'avons pas eu d'autre alternative que d'utiliser une approche statistique de normalisation globale. Cette approche globale considère que les différentes intensités des ions sont toutes reliées par un facteur constant entre les spectres. Il faut donc remettre toutes les intensités à la même échelle en les divisant une à une par un même coefficient. Pour cela, la normalisation par la médiane part du principe que, en moyenne, le nombre de protéines surexprimées est à peu près identique à celui des protéines sous-exprimées[239]. Elle considère également que le nombre de protéines dont l'abondance varie est faible par rapport au nombre total de protéines. Chaque spectre a donc été normalisé en divisant l'intensité de chaque ion par la médiane des intensités de tous les ions présents sur le spectre.

IV.3.2 Normalisations intra-échantillons

Une fois les spectres normalisés par la méthode de la médiane, les échantillons deviennent davantage comparables entre eux. Cependant, d'autres étapes sont nécessaires pour réduire l'influence du bruit et faire ressortir l'information biologique pertinente. Ces facteurs sont propres aux données de spectrométrie de masse. En premier lieu, il s'agit de la différence d'ordre de grandeur existant entre les abondances relatives des différentes espèces. Certaines protéines auront une abondance moyenne très faible par rapport à d'autres protéines très abondantes. Cependant, d'un point de vue biologique, les espèces abondantes ne sont pas obligatoirement plus intéressantes que celles très peu abondantes. Il existe également une légère fluctuation de l'abondance de certaines espèces dans des conditions expérimentales identiques (variabilité inter-individuelle). C'est ce que l'on résume sous le terme de variation biologique non-induite. Au final, les données de spectrométrie de masse sont sujettes à un bruit hétéroscédastique, ce qui signifie que l'écart-type des intensités augmente avec la valeur de l'intensité[240]. Autrement dit, le bruit n'est pas constant mais varie avec l'intensité du signal.

Il est donc nécessaire de réaliser, en plus de la normalisation par la médiane, plusieurs étapes de normalisation intra-échantillons. Elles se répartissent en trois classes : la transformation, le centrage et le redimensionnement des données.

Transformation :

La transformation est une étape de conversion non linéaire des données. Dans notre cas, nous avons choisi de remplacer chaque valeur d'intensité par son logarithme décimal (log_{10}) afin de corriger les phénomènes d'hétéroscédasticité et rendre la distribution des intensités plus symétrique[241].

Centrage :

Le centrage des données est un des traitements les plus couramment appliqués aux données spectrales. Il vise à répartir symétriquement l'intensité de chaque variable à travers les échantillons non plus autour de leur moyenne mais autour de 0. Cette méthode permet donc de réduire le décalage qui existe entre les protéines peu et très abondantes. Elle permet de se focaliser sur les différences existant entre les échantillons et non pas sur les similitudes[242].

Redimensionnement :

Le redimensionnement des données consiste à diviser chaque variable par un facteur unique, ce qui permet de réduire leur magnitude. Nous avons choisi d'utiliser la méthode de Pareto qui utilise la racine carrée de l'écart-type de la variable comme coefficient de redimensionnement[243]. A l'issue de cette transformation, les variances sont différentes d'une variable à l'autre mais la gamme de variance dans chaque spectre est largement réduite par rapport aux données initiales. Ainsi, l'importance relative des variables très intenses est-elle diminuée par rapport à celle des peu intenses. Cette méthode présente l'avantage de conserver au maximum la structure initiale des données, contrairement à d'autres méthodes de redimensionnement trouvées dans la littérature[244].

L'effet de chacune de ces étapes de normalisation a été exploré en utilisant une représentation en boîtes à moustache (*box plot*) des caractéristiques des variables. La figure 75 présente ces résultats sur 50 variables choisies aléatoirement parmi les 16 237 variables présentes dans la matrice *X*.

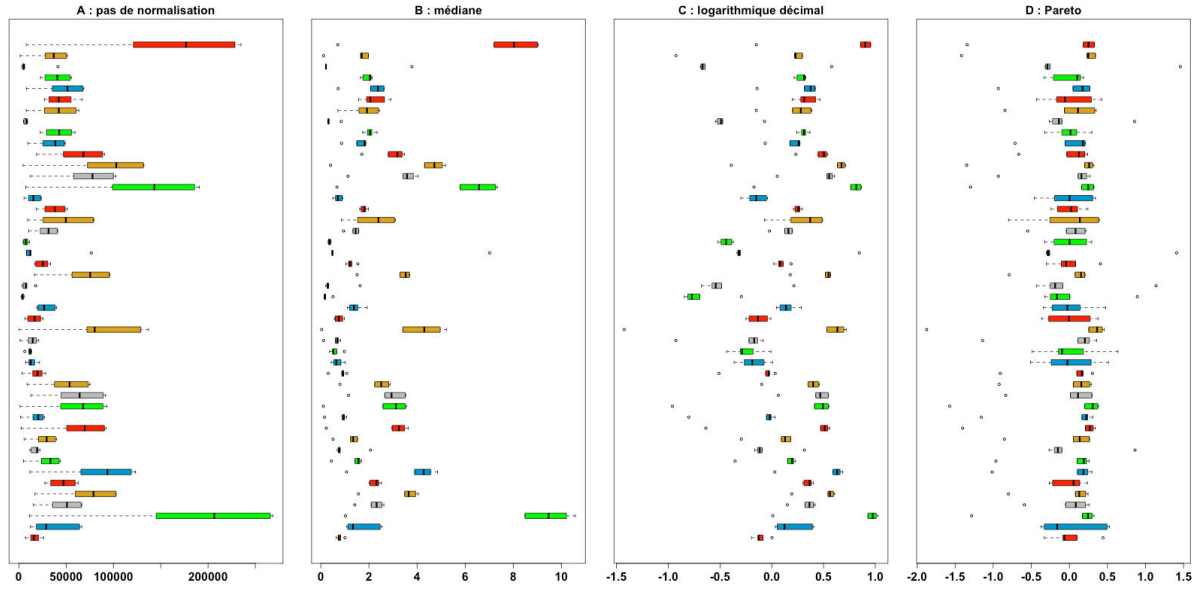


Figure 75 : boîtes à moustaches ou *box plot* résumant les caractéristiques à chacune des étapes de normalisation de 50 variables sélectionnées aléatoirement parmi les 16 237 variables de la matrice *X*. L'intensité des variables est représentée sur l'axe horizontal. A = pas de normalisation, B = normalisation par la médiane, C = transformation logarithmique et D = redimensionnement de Pareto.

Nous pouvons ainsi observer qu'au fil des étapes les variables tendent à se rapprocher d'une distribution centrée réduite. Ceci est confirmé en représentant la densité de probabilité de la distribution de l'intensité des variables au sein d'un échantillon avant et après les étapes de normalisation (figure 76).

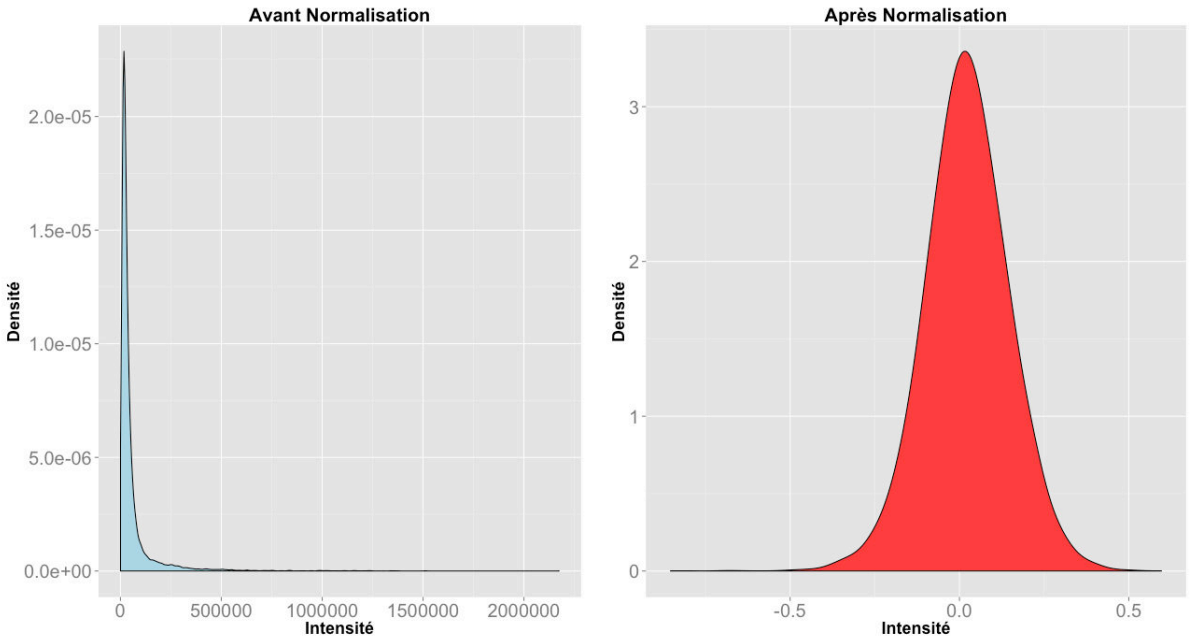


Figure 76 : estimation par noyau de la densité de probabilité de l'intensité des variables de la matrice *X* avant (gauche) et après (droite) les étapes de normalisation.

Au final, la distribution gaussienne des intensités des variables au sein des différents échantillons nous permettra d'utiliser différentes approches chimiométriques d'analyse des données en vue d'explorer les différences entre plusieurs groupes d'échantillons et ainsi en extraire des composés discriminants.

V. Approches chimiométriques pour l'analyse des données

V.1 Généralités

Le terme chimiométrie a été introduit pour la première fois par Svante Wold en 1974[245]. Il définit une réelle discipline scientifique dont la vocation est de sélectionner les procédures optimales pour acquérir des données de nature chimique et d'en extraire un maximum d'informations pertinentes. La chimiométrie fait donc appel à diverses méthodes issues de disciplines quantitatives : mathématiques appliquées, statistiques multivariées et informatique. Ces approches permettent d'explorer de grands volumes de données comprenant des mesures de plusieurs variables sur plusieurs échantillons à plusieurs instants. Dans notre cas, la complexité et la dimension des données acquises en spectrométrie de masse rendent archaïques les méthodes classiques univariées de visualisation des données. Des approches chimiométriques plus sophistiquées permettent d'explorer ces données dans leur globalité. Dans cette approche, il est primordial de préserver l'aspect combinatoire du code histone et en tenant compte des effets conjoints des variables, c'est-à-dire de la façon dont elles interagissent les unes avec les autres. Les approches statistiques multivariées utilisées en chimiométrie, à l'inverse des méthodes classiques univariées qui ne prennent en compte que quelques variables à la fois, permettent de considérer l'ensemble des variables. Elles constitueront ainsi une unité que l'on appelle une forme ou un motif. Les méthodes statistiques multivariées permettent d'identifier des formes à partir des données brutes et attribuent chaque forme à une catégorie d'échantillons. Ces formes identifiées sont ensuite comparées entre elles afin d'extraire les variables responsables de leurs similitudes et de leurs différences. C'est ce que l'on appelle la reconnaissance de formes ou *pattern recognition*,

branche de l'apprentissage automatique. Ces méthodes seront donc particulièrement utiles dans le cadre de notre approche histonomique globale puisqu'elles nous permettront de comparer les profils d'histones d'échantillons sains *versus* ceux d'échantillons exposés à des xénobiotiques que l'on soupçonne de perturber la régulation épigénétique. Elles répondent parfaitement à notre volonté de conserver l'information combinatoire du code histone et de considérer l'ensemble des variables comme une entité.

Les méthodes de reconnaissance de formes sont regroupées en deux classes distinctes : les méthodes non supervisées et les méthodes supervisées. Le choix de la méthode dépend du type d'information recherché. Les méthodes non supervisées sont des méthodes descriptives tandis que les méthodes supervisées sont prédictives. Nous détaillerons au cours de ce chapitre les raisons pour lesquelles nous avons employé ces méthodes ainsi que la nature de l'information que nous avons pu extraire de chacune d'elles. Les détails concernant le principe mathématique de ces méthodes ainsi que l'évaluation et l'interprétation des modèles générés sont présentés en partie expérimentale.

V.2 Méthodes non supervisées

Les méthodes non supervisées sont utilisées pour analyser un ensemble de variables sans attendre de réels résultats quantifiables. Ce sont des méthodes descriptives qui explorent les données de façon aveugle sans aucun *a priori* sur la nature des échantillons, considérés comme analogues. Elles nous permettent d'explorer la variabilité naturelle qui existe entre tous les échantillons et de révéler des structures à l'intérieur des matrices de variables. En résumé, elles mettent en évidence les tendances naturelles de regroupement qui peuvent exister entre les échantillons. Pour ce faire, les méthodes non supervisées utilisent principalement des représentations graphiques. Leurs résultats sont donc évalués visuellement.

V.2.1 Classification ascendante hiérarchique

La classification ascendante hiérarchique (CAH) est la première méthode non supervisée que nous avons mise en œuvre pour explorer les tendances naturelles de

regroupement entre nos échantillons. Sans indiquer la nature des échantillons (exposés ou non) cette méthode permet de les fractionner en groupes naturels ou *clusters*. L'approche que nous avons utilisée est dite agglomérative. Chaque échantillon constitue à la base son propre groupe, puis les groupes sont fusionnés deux à deux au fur et à mesure que l'on remonte dans la hiérarchie. Pour décider quels échantillons doivent être agglomérés ensemble, nous avons utilisé la méthode de Ward[246] comme critère de classification. La représentation en dendrogramme utilise la distance Euclidienne comme mesure de la dissimilarité entre les groupes de variables. Ainsi, plus les branches du dendrogramme sont longues, plus la distance Euclidienne est grande et plus les groupes sont différents. Le regroupement hiérarchique est donc une méthode intéressante pour relier le comportement global des variables avec la nature des échantillons. Cependant, au regard du très grand nombre de variables, elle trouve vite ses limites lorsqu'il s'agit d'extraire celles qui sont principalement responsables de la formation des *clusters*.

V.2.2 Analyse en composantes principales

V.2.2.1 Principe

L'analyse en composantes principales (ACP) est probablement la méthode non supervisée la plus populaire pour l'analyse multivariée de données biologiques. C'est la méthode exploratoire de choix lorsque l'on dispose de données volumineuses. Elle permet de réduire la dimension des données en les projetant dans un espace de plus faible dimension. Pour cela, son principe de base est de décomposer la matrice des variables *X* en une combinaison de la matrice des *scores T*, la matrice des *loadings P* et la matrice des résidus *E* (figure 77) de façon à obtenir la relation suivante : $\boldsymbol{X} = \boldsymbol{T} * \boldsymbol{P} + \boldsymbol{E}$.

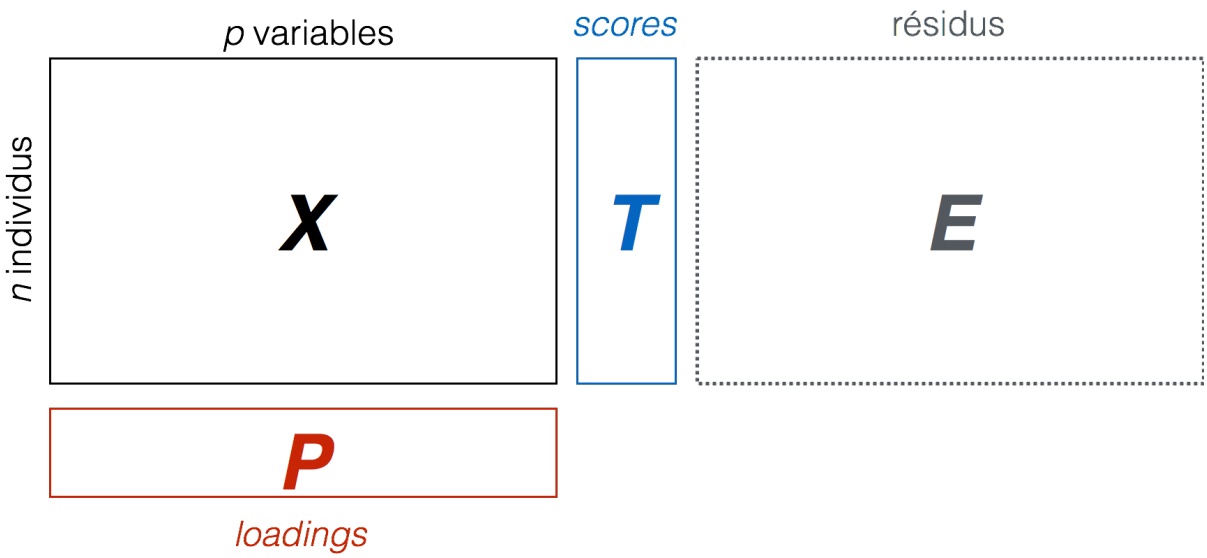


Figure 77 : décomposition matricielle de la matrice *X* effectuée lors d'une analyse en composantes principales.

L'ACP permet de révéler des structures cachées à l'intérieur des données en construisant des composantes principales (PC). Ces composantes sont des directions maximales de dispersion dans l'espace qui décrivent les sources de variance existant entre les échantillons. Elles correspondent en réalité aux vecteurs propres de la matrice des variances-covariances. Ces vecteurs propres sont calculés successivement afin de minimiser l'erreur résiduelle à chaque étape, et chaque vecteur propre consécutif représente un maximum de variation dans le jeu de données[247].

Chaque échantillon est projeté dans un plan tridimensionnel (*score plot*), et nous n'avons qu'à observer la répartition naturelle des échantillons dans cet espace pour repérer des tendances de regroupement, ou à l'inverse des directions de dispersion, voire des points qui ont un comportement totalement différent du reste des échantillons et que l'on qualifiera d'aberrants (*outliers)*. L'ACP nous permet de déterminer en un coup d'œil s'il existe des différences entre les profils d'histones des différents types d'échantillons et si elles sont reliées à la condition expérimentale étudiée, à savoir une exposition à un xénobiotique. Dans un second temps, chacune des variables peut également être projetée dans le même espace défini par les composantes principales (*loading plot*) afin d'identifier les principales variables responsables de la répartition géographique des échantillons dans le plan. L'ACP est donc très utile dans le cas d'une approche globale, mais ne suffit pas à établir un profil d'histones caractéristique d'une condition, profil qui

nous permettrait par la suite de classer les échantillons en fonction de leur code histone.

V.2.2.2 Choix du nombre de composantes

D'un point de vue théorique, il existe autant de composantes principales que de direction possible de dispersion des échantillons. Autrement dit, il existe autant de composantes principales que d'échantillons étudiés lors d'une ACP. En pratique, le choix du nombre de composantes est très important et conditionne la qualité du modèle statistique qui en découle. Pour faire ce choix, deux paramètres doivent être pris en compte : le degré d'ajustement du modèle aux données et son pouvoir prédictif. Le degré d'ajustement du modèle peut être évalué quantitativement à l'aide du paramètre R^2X qui traduit la part de variation expliquée mathématiquement par le modèle. Plus R^2X est élevé, plus le modèle capte une part importante de la variation existante dans le jeu de données d'apprentissage. Cependant, un paramètre R^2X élevé pris en compte seul n'est pas garant de la qualité d'un modèle statistique. En effet, le modèle peut capter du bruit ou une source de variation non pertinente comme lors de la présence d'individus aberrants (*outliers*). Il peut ainsi atteindre arbitrairement la valeur maximale de 1, soit 100% de la variation expliquée par le modèle. Il existe un paramètre plus significatif que le degré d'ajustement. Il s'agit du pouvoir prédictif d'un modèle. Il est mesuré à l'aide du paramètre quantitatif Q^2X qui représente le pouvoir prédictif du modèle construit.

Les paramètres R^2X et Q^2X ont des comportements différents en fonction du degré de complexité du modèle, c'est-à-dire en fonction du nombre de composantes sélectionnées. Le paramètre R^2X tend vers l'ajustement parfait, soit la valeur 1, lorsque la complexité du modèle augmente. Le paramètre Q^2X quant à lui ne tend pas obligatoirement vers 1 et peut décroître lorsque le modèle devient trop complexe. Il faut donc trouver le meilleur équilibre entre la complexité du modèle et la part de variation prédite. Ce compromis se situe au niveau du plateau de Q^2 comme illustré sur la figure 78.

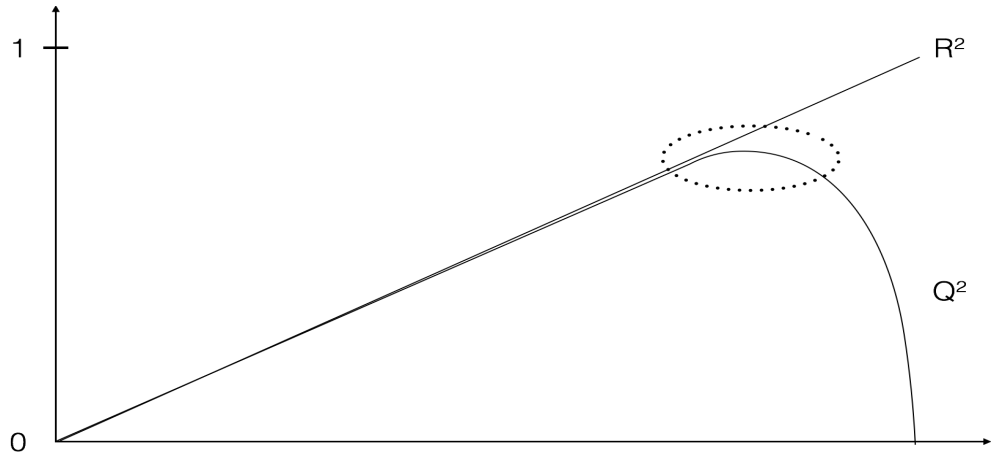


Figure 78 : évolution des paramètres R^2 et Q^2 en fonction du nombre de composantes sélectionnées dans le modèle. L'axe vertical représente la part de variation expliquée ou prédite. L'axe horizontal représente la complexité du modèle. La valeur de Q^2 atteint un plateau entouré ici en pointillés.

Dans notre cas nous avons utilisé la procédure de validation croisée[248] (*cross-validation,* CV) sur le R^2 et Q^2. Cette procédure de validation interne vise à trouver la dimension optimale d'un modèle afin qu'il présente les meilleures performances. L'idée de base de la validation croisée consiste à exclure une partie des données initiales lors de la construction du modèle puis de se servir du modèle pour prédire ces données. Les valeurs prédites sont comparées avec les valeurs réelles afin d'évaluer la performance du modèle. Cette procédure est répétée plusieurs fois jusqu'à ce que chaque individu ait été exclu une seule fois. Au final, les carrés des différences observées entres les valeurs prédites et les valeurs réelles sont additionnées afin de calculer le PRESS (*Predictive Residual Sum of Squares*), mesure du pouvoir prédictif. Lorsqu'on augmente la complexité du modèle, chaque composante est considérée comme significative si le PRESS divisé par la somme des carrées résiduels (*residual sum of squares*, SS) de la composante précédente est inférieur à 1.

V.2.2.3 Interprétation des résultats

Une ACP fournit principalement des représentations graphiques qui permettent d'interpréter les résultats. A partir du modèle généré et validé, deux types de représentation sont utiles pour visualiser la variabilité naturelle qui existe à l'intérieur d'un jeu de données (figure 79). La première représentation graphique est appelée *scores plot*. Elle correspond à la projection des échantillons dans

l'espace défini par les composantes principales. Elle permet donc de visualiser les tendances de regroupement des échantillons en fonction de leurs similitudes et de leurs différences. En parallèle, chaque variable peut être projetée dans ce même espace sur un *loadings plot*, qui permet, en le superposant au *scores plot*, d'identifier les variables responsables des regroupements observés.

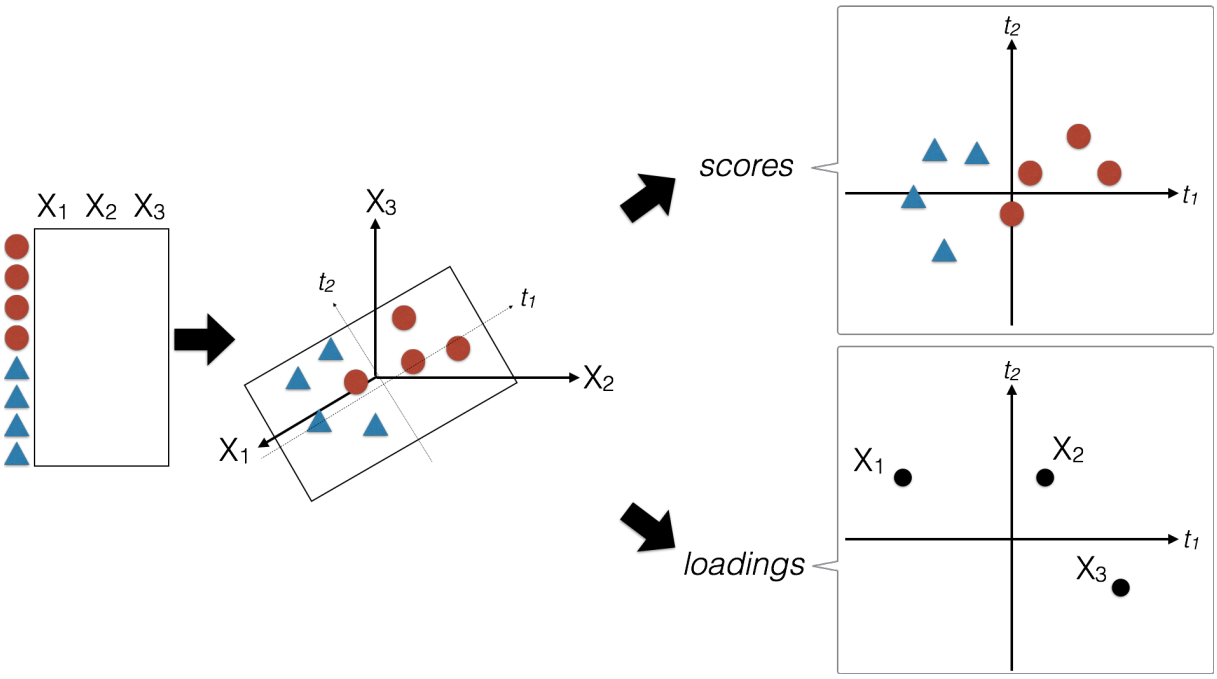


Figure 79 : principe de construction et d'interprétation des résultats d'une analyse en composantes principales.

V.3 Méthodes supervisées

L'objectif final l'approche globale mise au point étant de discriminer les échantillons sur la base de leur profil d'histones, les méthodes supervisées se sont imposées comme étant des méthodes de choix. Elles considèrent les échantillons non plus de manière naïve mais en fonction d'une réponse observée. Elles englobent des méthodes de régression et de classification en fonction du problème posé. Les premières permettent de prédire une valeur numérique à l'image d'une droite de régression linéaire, tandis que les secondes permettent de prédire l'appartenance à une classe d'échantillons. Pour notre part nous avons utilisé les méthodes de classification. A partir d'un jeu de données d'apprentissage (*training set*), chaque échantillon est étiqueté en fonction de son appartenance à une classe d'échantillons (exposés ou non exposés) afin de constituer la matrice des classes *Y*.

Les méthodes de classification apprennent à reconnaître à partir du *training set* les critères qui ont permis de classer les échantillons dans chacun des groupes déterminés.

V.3.1 Analyse discriminante PLS

V.3.1.1 Principe

L'analyse discriminante par la méthode des moindres carrés partiels (*Partial Least Squares/Projection to Latent Structures-Discriminant Analysis*, PLS-DA) est une méthode de classification dont le principe est de maximiser la covariance entre la matrice *X* des variables et la matrice *Y* des classes[249] (figure 80).

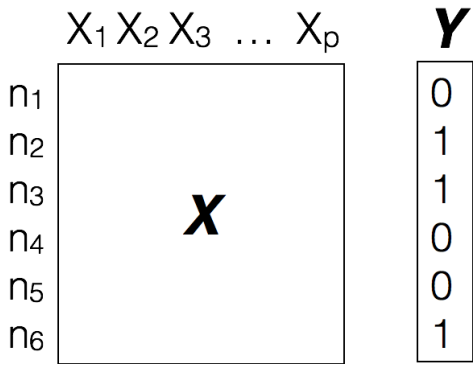


Figure 80 : lors d'une analyse PLS-DA, la matrice *Y* est créee afin de labelliser chacun des échantillons présents dans la matrice *X* comme appartenant à une classe (0 ou 1 dans le cas d'une comparaison entre deux classes).

Pour cela, elle réalise une rotation de la projection dans l'espace afin d'obtenir des variables latentes qui se concentrent sur la discrimination des classes[250]. Contrairement aux analyses non supervisées, elle prend en compte l'appartenance des échantillons à une classe dès les premières étapes de la construction du modèle. Le modèle généré permettra donc d'obtenir une séparation maximale entre les classes sur la base des variables contenues dans la matrice *X*, puis d'extraire les variables responsables de cette discrimination. Cette analyse discriminante requière une certaine homogénéité à l'intérieur de chaque classe et peut modéliser entre 2 et 4 classes différentes. Au-delà, la discrimination devient hasardeuse.

V.3.1.2 Choix du nombre de composantes : validation croisée

La validation croisée se fait selon la même procédure que celle décrite précédemment pour les modèles ACP. Elle permet de déterminer le nombre nécessaire de composantes pour capturer suffisamment de variation sans intégrer de bruit au modèle. Les paramètres R^2(cum) et Q^2(cum) obtenus en additionnant respectivement les valeurs de R^2 et Q^2 de chaque composante retenue permettent d'estimer le pouvoir prédictif du modèle. Ces valeurs doivent ainsi être les plus élevées et les plus proches possibles l'une de l'autre pour garantir la fiabilité du modèle. Il survient parfois un phénomène de surapprentissage qui signifie que le modèle capte de l'information dans les données qui n'est en fait que du bruit et qui n'offrira aucun pouvoir prédictif. C'est pour éviter cela que les modèles PLS-DA doivent être soigneusement validés avant d'en tirer une quelconque interprétation biologique.

V.3.1.3 Validation des modèles

CV-ANOVA :

Une des limites du paramètre Q^2 fourni par la validation croisée est qu'il n'évalue que le pouvoir prédictif d'un modèle, mais ne fournit aucune information sur la significativité statistique du pouvoir prédictif estimé. Une des manières d'obtenir cette information est d'utiliser le test CV-ANOVA (*ANalysis Of VAriance testing of Cross-Validated predictive residuals*). Ce test est un véritable outil diagnostique qui évalue la fiabilité d'un modèle PLS-DA. Il consiste à réaliser un test d'hypothèse ANOVA sur les résidus obtenus après validation croisée du modèle[251]. Plus simplement, il teste si les résidus issus de la prédiction par le modèle PLS-DA sont significativement inférieurs à la simple variation moyenne. Une *p-value* faible indique donc que les résidus prédits sont inférieurs aux résidus moyens et attestent de la significativité du modèle.

Test de permutation :

Le test de permutation est un des moyens utilisés pour valider un modèle supervisé. Il est plus complexe que le test CV-ANOVA et requiert davantage de

temps de calcul, en particulier si le modèle est complexe. En partant des données d'apprentissage initiales, la matrice *X* est laissée intacte, tandis que les valeurs de la matrice *Y* sont réarrangées entre elles de façon aléatoire pour apparaître dans un ordre différent. Un modèle PLS-DA est alors généré à partir de la matrice *Y* permutée, et les valeurs de R^2 et Q^2 sont calculées par validation croisée. Cette procédure est répétée *x* fois (entre 25 et 999 fois) aboutissant à la génération de *x* modèles PLS-DA permutés. La distribution des valeurs de R^2 et Q^2 à travers ces modèles permutés est ensuite comparée aux valeurs de R^2 et Q^2 du modèle réel. Pour que le modèle soit considéré comme valide, il faut que ses valeurs de R^2 et Q^2 soit supérieures aux valeurs d'intersection entre l'axe des ordonnées et les droites de régression linéaire de R^2 et Q^2 issues des modèles permutés.

Validation externe :

Les méthodes de validation mathématiques telles que la validation croisée et le test de permutation sont des méthodes de validations internes. Elles fournissent une idée raisonnable du pouvoir prédictif d'un modèle PLS-DA. Cependant, il existe une façon beaucoup plus rigoureuse et drastique de valider un modèle PLS-DA. Il s'agit de la validation externe à partir de données test. Cette procédure consiste à utiliser un jeu de données indépendant jamais utilisé auparavant lors de la construction du modèle, et à prédire l'appartenance des échantillons à l'une des classes. Le pouvoir prédictif du modèle se mesurera alors par le pourcentage de prédictions justes. Ces données test doivent être représentatives des données d'apprentissage. La validation externe met donc le modèle à l'épreuve de la réalité et rend compte de sa capacité éventuelle à être utilisé en pratique.

V.3.1.4 Interprétation des résultats

L'interprétation des résultats d'une analyse PLS-DA repose en premier lieu sur la validation du modèle. En effet, les représentations graphiques de type *scores plot* sont en réalité artéfactuelles et la séparation que l'on peut y observer ne reflète pas obligatoirement la différence qui existe réellement entre les classes. En revanche, si le modèle est correctement validé, la séparation observée est significative et représente une réelle différence dans les données. L'étape suivante consistera donc à extraire les variables responsables de la discrimination entre les

classes. Pour cela, il existe un paramètre qui mesure l'importance de la variable dans la projection, c'est-à-dire sa contribution relative à la séparation des classes. Ce paramètre est appelé score VIP (*Variable Importance in the Projection*). La somme des carrées de tous les scores VIP d'un modèle étant égale au nombre de variables présentes dans la matrice *X*, le score VIP moyen est égal à 1. Ainsi, il est possible de comparer les scores VIP entre eux, et un score supérieur à 1 traduit une contribution significative de la variable dans la discrimination des classes.

V.3.2 Analyse discriminante OPLS

V.3.2.1 Principe

L'analyse discriminante OPLS[252] (*Orthogonal Projection to Latent Structures-Discriminant Analysis*, OPLS-DA) est une modification de la méthode PLS-DA qui sépare la variabilité présente dans la matrice *X* en deux parties : une partie prédictive qui est linéairement reliée à la matrice *Y*, et une partie orthogonale qui n'est pas reliée à la matrice *Y*. C'est une méthode particulièrement adaptée à la discrimination de deux classes différentes. Les modèles générés possèdent les mêmes propriétés que les modèles PLS-DA, à la différence que l'information prédictive recherchée est concentrée sur la première composante, tandis que toute la variabilité présente dans la matrice *X* qui n'est pas reliée à la discrimination entre les deux classes est placée sur la ou les composante(s) orthogonale(s). Cette partition de la variabilité améliore sensiblement la transparence des modèles et simplifie leur interprétation. C'est donc la méthode de choix pour comparer des échantillons exposés ou non à un xénobiotique et pour extraire uniquement l'information discriminante reliée à l'exposition.

V.3.2.2 Choix du nombre de composantes et validation des modèles

Les modèles OPLS-DA sont dérivés des modèles PLS-DA et possèdent les mêmes caractéristiques et le même pouvoir prédictif. Ainsi, tous les paramètres de sélection des composantes et de validation sont identiques à ceux des modèles PLS-DA, hormis le test de permutation qu'il n'est pas possible d'effectuer pour un modèle OPLS-DA.

V.3.2.3 Interprétation des résultats

L'interprétation des résultats et l'extraction des variables discriminantes à partir des modèles OPLS-DA se font de la même manière que pour les modèles PLS-DA. Les scores VIP sont un bon reflet de l'importance de chaque variable dans la discrimination des classes. Il existe cependant une représentation graphique propre aux logiciels Umetrics qui permet d'extraire plus facilement les variables discriminantes des modèles OPLS-DA. Il s'agit du S-plot^{TM}, qui fournit une visualisation du poids des variables sur la composante prédictive. Le S-plot^{TM} représente la covariance et la corrélation existantes entre les variables et les scores prédictifs[253]. Lorsque les données ont été préalablement centrées et redimensionnées, cette représentation prend souvent la forme de la lettre S. Les variables éloignées horizontalement des extrémités du S ont à la fois une forte influence dans la discrimination et une excellente fiabilité. Ce sont donc les variables les plus intéressantes lorsqu'on le cherche des marqueurs sur- ou sous-exprimés en fonction de la condition expérimentale étudiée.

V.4 Conclusion

Les méthodes chimiométriques choisies permettent d'explorer des données volumineuses de manière simple et intuitive tout en conservant les relations complexes existant entre les différentes variables. Les méthodes non supervisées offrent une vue d'ensemble de la variabilité présente et fournissent un premier aperçu des variables potentiellement discriminantes. Les méthodes supervisées permettent, quant à elles, de séparer les échantillons en fonction de leur appartenance à une classe et d'extraire des listes de variables contribuant significativement à la discrimination entre classes. L'étape suivante consiste donc à valider statistiquement chacune des variables sélectionnées puis à les identifier afin de révéler de potentiels marqueurs spécifiques d'exposition à un xénobiotique.

VI. Validation et interprétation des résultats

La dernière étape de notre stratégie globale vise à valider les résultats obtenus par les approches multivariées et d'identifier les variables sélectionnées. Pour cela, plusieurs étapes sont nécessaires.

VI.1 Coefficient de variation

Pour valider les résultats obtenus lors des analyses statistiques multivariées, la première étape consiste à s'assurer que la variation d'abondance relative des protéines discriminantes observée entre les groupes d'échantillons n'est pas due à la variabilité analytique du système au cours du temps. Pour cela, nous avons utilisé l'ensemble des échantillons QC injectés tout au long de l'analyse. En effet, les échantillons QC sont par nature représentatifs de la composition en histones de la totalité des échantillons de l'étude. Nous avons estimé la variabilité analytique pour chaque ion sélectionné lors des analyses multivariées en calculant le coefficient de variation (CV) de son intensité à travers l'ensemble des échantillons QC. Plus le coefficient de variation est faible, plus la mesure de l'intensité pour un ion est reproductible. La FDA (*Food and Drug Administration*) a fixé la valeur maximale pour les CV à 20% dans le cadre d'une recherche de biomarqueurs[254]. Cependant, compte-tenu du type d'étude exploratoire que nous avons effectué ainsi que de la nature des molécules sur lesquelles nous avons travaillé, nous nous sommes fixés ce seuil d'acceptabilité à 30%. Ainsi, toutes les variables pour lesquelles les CV des intensité à travers les échantillons QC excédaient 30% ont été éliminées, une telle variation rendant difficile une attribution à un phénomène biologique plutôt qu'analytique.

VI.2 Tests statistiques univariés

L'approche histonomique globale a pour objectif final de révéler des formes d'histones dont l'abondance relative varie en réponse à une exposition à un xénobiotique. Les analyses statistiques multivariées nous permettent d'extraire les ions discriminants, mais ne fournissent pas de garantie quant à la significativité des différences observées entre les deux populations. Il est donc nécessaire de

comparer un paramètre statistique entre chaque population afin de déterminer si les différences observées sont dues au phénomène étudié ou au hasard. Les données étant normalisées, nous avons choisi d'utiliser un test d'hypothèse paramétrique, à savoir le test t de Welch. Ce test statistique univarié nous permet de considérer chaque variable indépendamment et de tester la significativité des différences observées entre les échantillons témoins et exposés pour cette variable. Ce test est une adaptation du test t de Student qui peut être utilisé lorsque la distribution des populations est normale mais n'exige pas que les deux populations aient le même écart-type[255]. Tout d'abord, l'hypothèse nulle H_0 est posée et affirme que les différences observées entre les moyennes des deux populations sont dues au hasard. Le seuil de probabilité α pour rejeter l'hypothèse nulle (erreur de type 1) est fixé à 1%. Le test d'hypothèse est ensuite effectué sur chacune des moyennes des variables sélectionnées et nous fournit la probabilité de rejeter l'hypothèse nulle que l'on note *p-value*. Si cette *p-value* est en dessous du seuil α, c'est-à-dire inférieure à 0,01, l'hypothèse nulle est rejetée et les différences observées sont considérées comme significatives.

VI.3 Test d'hypothèses multiples

En considérant le nombre important de variables discriminantes sélectionnées lors des analyses multivariées, il y aura autant de tests univariés effectués en parallèle. En statistique, plus le nombre de tests d'hypothèses augmente, plus la probabilité due au hasard de rejeter à tort l'hypothèse nulle (erreur de type 1) augmente. Ce phénomène conduit à un nombre conséquent de faux positifs et est résumé sous le terme de problème des comparaisons multiples. Nous avons ainsi ajusté les *p-value* fournies par chaque test effectué en parallèle afin d'éviter de fausser les résultats obtenus. Pour cela, nous avons utilisé le taux de fausse découverte ou FDR (*False Discovery Rate*). Le test FDR est une méthode de correction classiquement utilisée en cas multiplicité des tests d'hypothèse[256]. Il permet d'estimer le taux de faux positifs parmi un grand nombre de variables en calculant des *q-value* correspondant aux *p-value* corrigées pour chacun des tests t effectués. Pour chaque variable, la *q-value* peut être comprise comme étant la proportion de faux positifs attendue lorsque cette variable est considérée comme significativement différente entre les deux populations[257]. Ainsi, une *q-value* égale

à 0,01 signifie que 1% de tous les tests seront des faux positifs tandis qu'une *q-value* égale à 0,01 signifie que 1% de tous les tests significatifs seront des faux positifs. Dans notre cas nous avons fixé le seuil d'acceptabilitié à 1%, ce qui nous permet d'avoir confiance en la significativité des différences observées et de nous assurer que l'abondance des variables discriminantes est réellement différente entre les échantillons témoins et les échantillons exposés à un xénobiotique.

VI.4 Calcul des ratios d'abondance

Les ratios d'abondance (*Fold Change*, FC) sont fréquemment utilisés pour évaluer la magnitude des différences d'abondance observées entre deux populations. Pour calculer ces ratios, nous avons utilisé les moyennes des intensités de chaque variable sélectionnée dans le groupe témoin et le groupe exposé. La méthode de calcul utilisée est détaillée en partie expérimentale. Nous avons utilisé le logarithme binaire des ratios d'abondance pour faciliter l'interprétation des variables sur- ou sous-exprimées. Une valeur négative du logarithme binaire du ratio d'abondance signifiera ainsi que la variable est sous-exprimée lors d'une exposition au xénobiotique étudié. A l'inverse, si cette valeur est positive cela signifiera que la variable en question est surexprimée lors de l'exposition. Nous ne nous sommes pas fixé de seuil minimal concernant la magnitude du changement d'abondance car le moindre changement, aussi faible soit-il, doit être pris en considération dans un contexte de perturbation épigénétique.

VI.5 Identification des variables validées

VI.5.1 Déconvolution MaxEnt1

L'algorithme MaxEnt1 intégré au logiciel MassLynx se fonde sur le principe d'entropie maximale[258] pour produire des spectres de masses moléculaires moyennes à partir de séries d'ions multichargés sur les spectres ES (figure 81). Cet algorithme trouve le spectre de masse moléculaire le plus simple, c'est-à-dire le spectre d'entropie maximale, qui correspond aux rapports *m*/*z* observés sur les spectres originaux. Il fonctionne de manière itérative en prenant une première approximation du spectre de masse moléculaire, puis en y intégrant des

informations physico-chimiques préalablement programmées pour générer un spectre fictif. Il compare alors ce spectre fictif aux données originales et utilise la différence observée entre les deux pour affiner le processus et générer un nouveau spectre fictif. L'algorithme répète cette procédure un certain nombre de fois jusqu'à ce que la différence entre le spectre de déconvolution fictif et le spectre initial soit négligeable. Le spectre de déconvolution obtenu par l'algorithme MaxEnt1 conserve la forme gaussienne des pics et l'aire sous les pics est représentative de l'abondance relative de l'espèce considérée. La qualité du spectre déconvolué dépend de certains paramètres spécifiés qui sont propres à chaque spectre. Ces paramètres sont détaillés en partie expérimentale.

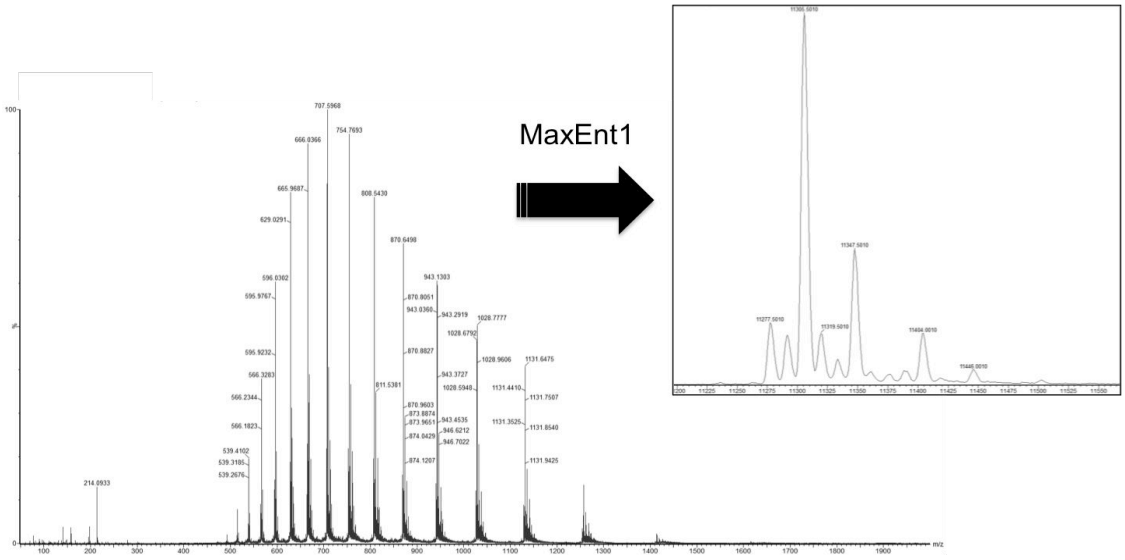


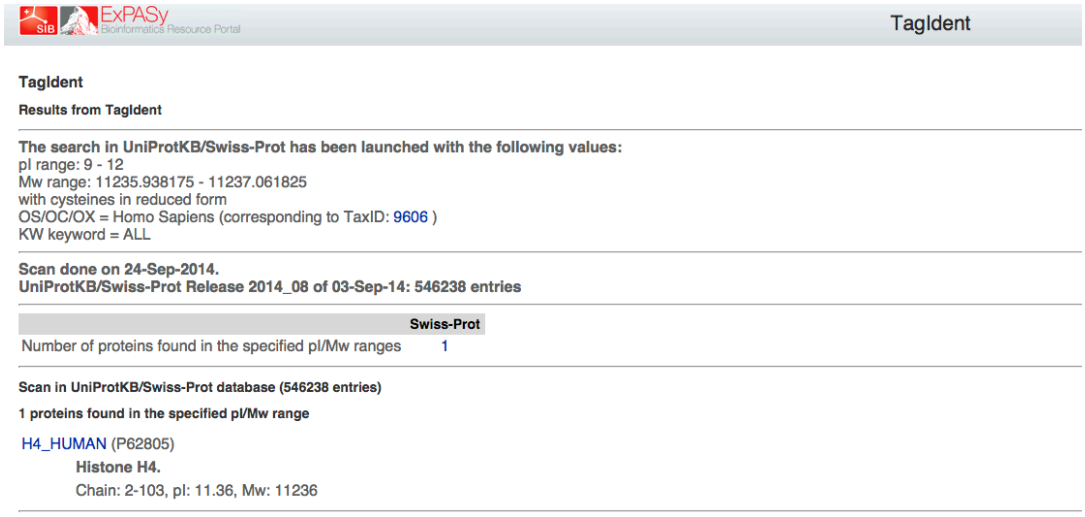
Figure 81 : exemple de la déconvolution par MaxEnt1 d'un spectre ESI de l'histone H4 extraite de cellules BeWo. Cette déconvolution permet d'obtenir les masses moléculaires des protéines entières à partir des rapports *m*/z sur un spectre ESI.

Pour identifier les variables sélectionnées lors des analyses statistiques, il suffit ainsi de les rechercher sur les spectres continuum à l'aide de leur temps de rétention et de leur rapport *m*/z, puis de repérer sur les spectres déconvolués les masses moléculaires des protéines entières auxquelles ces variables correspondent.

Cette démarche nous fournit donc les masses moléculaires moyennes des espèces discriminantes. Cependant, en considérant la diversité des formes d'histones présentes en mélange ainsi que la complexité des profils de modifications post-traductionnelles, la seule masse moléculaire ne suffit pas pour en déduire la forme d'histone dont il s'agit.

VI.5.2 Moteur de recherche TagIdent

Afin d'identifier le plus précisément possible à quelles formes d'histones correspondent les différentes variables discriminantes, nous avons choisi d'utiliser l'outil TagIdent disponible *via* le serveur ExPASy[259]. Il s'agit d'un moteur de recherche capable de créer des listes de protéines correspondant à certains critères spécifiés par l'utilisateur tels que l'organisme, la gamme de pI, la gamme de masse moléculaire attendue et l'erreur relative sur la mesure de masse. Il permet d'effectuer une recherche sur la base de données de séquences protéiques UniProtKB/Swiss-Prot[260] et propose une liste de protéines correspondant aux critères spécifiés (figure 82). Dans le cas de nos échantillons, les protéines présentes en mélange ont été obtenues par une extraction acide. Nous avons donc considéré que toutes étaient très basiques, avec un pI compris entre 9 et 12. L'erreur de masse relative de 0,5 Da a été calculée à partir des masses moyennes obtenues sur les spectres déconvolués. Les recherches ont donc été effectuées avec ces paramètres pour toutes les masses moléculaires moyennes correspondant aux variables discriminantes. L'outil TagIdent nous permet donc d'identifier avec une certaine confiance les protéines présentes en mélange à partir de leurs masses moléculaires moyennes.


ExPASy Bioinformatics Resource Portal — TagIdent

TagIdent

Results from TagIdent

The search in UniProtKB/Swiss-Prot has been launched with the following values:
pI range: 9 - 12
Mw range: 11235.938175 - 11237.061825
with cysteines in reduced form
OS/OC/OX = Homo Sapiens (corresponding to TaxID: 9606)
KW keyword = ALL

Scan done on 24-Sep-2014.
UniProtKB/Swiss-Prot Release 2014_08 of 03-Sep-14: 546238 entries

	Swiss-Prot
Number of proteins found in the specified pI/Mw ranges	1

Scan in UniProtKB/Swiss-Prot database (546238 entries)

1 proteins found in the specified pI/Mw range

H4_HUMAN (P62805)
Histone H4.
Chain: 2-103, pI: 11.36, Mw: 11236

Figure 82 : capture d'écran d'une recherche effectuée sur TagIdent. Les paramètres suivants ont été utilisés : Taxonomie = Homo Sapiens, gamme de pI = 9 – 12, masse moyenne observée sur le spectre déconvolué = 11236,5 Da, erreur relative sur la mesure de masse = 0,005%. Le moteur de recherche nous fournit un résultat unique correspondant ici à l'histone H4 humaine.

Partie 3

Application à deux cas d'exposition à des xénobiotiques

Chapitre I

Exposition à un inhibiteur HDAC : le butyrate de sodium

Chapitre II

Exposition à un agent toxique: le benzo[*a*]pyrène

I. Exposition à un inhibiteur HDAC : le butyrate de sodium

I.1 Introduction

L'acétylation est une modification post-traductionnelle très fréquente sur les histones. Elle intervient principalement sur les arginines et les lysines et est catalysée par une enzyme de la famille des histones acétyltransférases (HAT). La réaction inverse est catalysée par des enzymes de la famille des histones désacétylases (HDAC). Une acétylation des histones est classiquement associée à une diminution locale de l'interaction entre les histones et l'ADN et donc à une décondensation de la chromatine aboutissant à une augmentation de la transcription des gènes.

Le degré d'acétylation des histones ayant un impact direct sur le niveau de transcription des gènes, la perturbation de l'équilibre entre acétylation et désacétylation se révèle être impliquée dans le développement de certaines pathologies chroniques. C'est notamment le cas de nombreux cancers dans lesquels l'hypo- ou l'hyperacétylation des histones est directement corrélée à la sévérité ou au bon pronostic selon les tissus concernés[261]. Une stratégie thérapeutique en cancérologie consistant à maintenir cet équilibre a donc naturellement émergée. Les inhibiteurs HDAC (HDACI) sont ainsi considérés depuis plusieurs années comme des agents thérapeutiques dans la prévention et le traitement de certains cancers[262]. En inhibant la désacétylation des histones, les HDACI modulent l'expression de gènes cibles associés aux phénotypes malins[263]. Cependant, cette inhibition doit être sélective. L'hyperacétylation relative qui résulte d'un traitement par un HDACI non sélectif peut en effet affecter l'activité de certaines protéines autres que les histones et entraîner la survenue d'effets secondaires[264]. Plusieurs molécules chimiques ayant la capacité d'inhiber sélectivement certaines HDAC ont été testées cliniquement durant ces dix dernières années. Parmi elles, seules deux molécules se sont révélées avoir une balance bénéfice/risque positive et ont été approuvées par la FDA[265] : le vorinostat et la romidepsine. D'autres HDACI sont actuellement en phase d'essais cliniques et pourraient venir compléter cet arsenal thérapeutique[266].

En plus de leur implication dans le développement de certains cancers, de plus en plus d'études montrent que les HDAC sont également impliquées dans le développement de pathologies cardiaques[267] et neurodégénératives[268] dont la maladie de Parkinson[269] et la maladie d'Alzheimer[270].

Ainsi les HDACI représentent une classe à part entière d'agents thérapeutiques dont les applications ne cessent de s'étendre. Cependant, comme chaque substance active, les HDACI présentent certains effets indésirables qui surviennent de manière dose-dépendante. L'hyperacétylation induite par les HDACI doit être surveillée de près afin de suivre l'efficacité du traitement et d'adapter les doses aux effets recherchés. Dans cette optique, notre approche histonomique globale peut s'avérer être un outil précieux pour suivre la modulation du degré d'acétylation des histones induite par un traitement HDACI. De plus, l'exposition à un HDACI dont l'effet sur les histones est attendu nous permet de valider notre approche sur un cas réel d'exposition à un xénobiotique.

I.2 Exposition au butyrate de sodium

Pour mettre la stratégie mise au point à l'épreuve et prouver son utilité pour le suivi du degré d'acétylation des histones, nous avons choisi d'exposer des cellules BeWo (clone b30) au butyrate de sodium (BS). Le BS est un inhibiteur non compétitif et non sélectif des HDAC parmi les plus anciennement connus[271]. C'est un acide gras à chaîne courte (figure 83) produit par fermentation bactérienne aérobie des fibres alimentaires. Lorsqu'il est introduit dans le milieu de culture à des concentrations relativement faible (de l'ordre de 5 mM) il est capable d'induire une hyperacétylation des histones et dans certains cas de stopper le cycle cellulaire[272]. Il peut également avoir d'autres effets selon les types cellulaires, comme une modification de la morphologie des cellules ou un remodelage du cytosquelette[273].

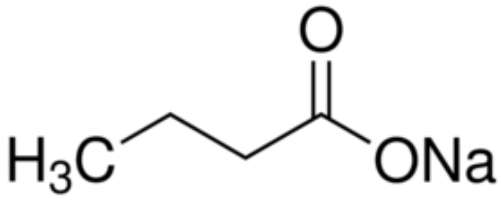


Figure 83 : formule topologique du butyrate de sodium.

Dans le but de vérifier si l'approche histonomique était capable de révéler un effet dose-réponse, nous avons exposé deux lots de 15 flasques chacun à deux doses de BS différentes. Le premier lot a été exposé au BS à 1 mM pendant 24 heures et le second au BS à 2,5 mM pendant 24 heures sur des cellules à confluence. Ces deux doses ont volontairement été choisies en dessous de 5 mM à partir de laquelle un effet apoptotique peut être observé. Un lot supplémentaire de 15 flasques n'a été exposé qu'au vecteur, c'est-à-dire au milieu de culture F-12K seul, afin de servir de groupe témoin.

I.3 Extraction et profilage LC-MS des histones

I.3.1 Dosage des protéines et contrôles

Les histones ont été extraites à partir des culots cellulaires selon le protocole d'extraction acide détaillé en partie expérimentale. Au total, 45 extraits histoniques ont été obtenus et tous ont été dosés par la méthode BCA. Les résultats du dosage pour les échantillons précipités par le TCA sont présentés dans le tableau 20. Les concentrations obtenues peuvent varier d'un échantillon à un autre, probablement à cause de la variation du nombre de cellules d'une flasque à une autre.

Tableau 20 : concentrations des différents mélanges d'histones extraits à partir de cellules BeWo non exposées ou exposées au butyrate de sodium à 1 ou 2,5 mM.

	Conc. Histones (µg/µL)
Témoin_1	0,79
Témoin_2	0,87
Témoin_3	1,10

Témoin_4	0,790
Témoin_5	0,80
Témoin_6	0,74
Témoin_7	0,78
Témoin_8	0,10
Témoin_9	0,88
Témoin_10	0,63
Témoin_11	0,78
Témoin_12	0,78
Témoin_13	0,75
Témoin_14	0,80
Témoin_15	0,81
SB 1mM_1	0,67
SB 1mM_2	0,71
SB 1mM_3	0,73
SB 1mM_4	0,68
SB 1mM_5	0,71
SB 1mM_6	0,62
SB 1mM_7	0,68
SB 1mM_8	0,64
SB 1mM_9	0,79
SB 1mM_10	0,64
SB 1mM_11	0,63
SB 1mM_12	0,68
SB 1mM_13	0,50
SB 1mM_14	0,59
SB 1mM_15	0,50
SB 2.5mM_1	0,71
SB 2.5mM_2	0,71
SB 2.5mM_3	0,89
SB 2.5mM_4	0,79
SB 2.5mM_5	0,67
SB 2.5mM_6	0,77
SB 2.5mM_7	0,58
SB 2.5mM_8	0,60
SB 2.5mM_9	0,71
SB 2.5mM_10	0,74
SB 2.5mM_11	0,68
SB 2.5mM_12	0,84
SB 2.5mM_13	0,61
SB 2.5mM_14	0,83
SB 2.5mM_15	0,80

Une quantité équivalente de chacun de ces extraits (environ 5 µg) a été déposée sur SDS-PAGE 13% (figure 84). Une fois les contrôles effectués, tous les

échantillons ont été dilués dans une solution aqueuse d'acide formique 0,05% (solvant A) afin d'obtenir une concentration unique de 0,3 µg/µL.

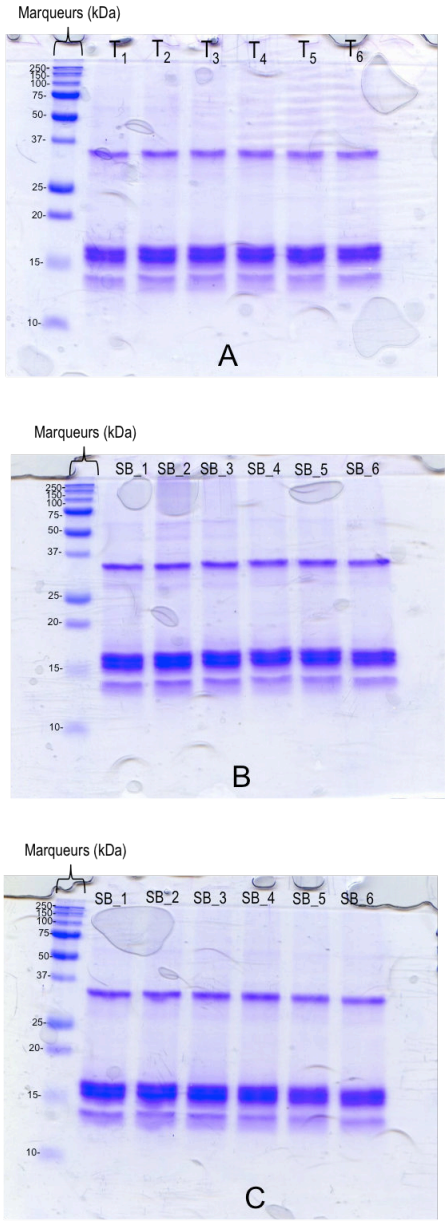


Figure 84 : SDS-PAGE 13% montrant les profils électrophorétiques d'extraits histoniques obtenus à partir de cellules BeWo non exposées (A) ou exposées au BS à 1 mM (B) ou 2,5 mM (C).

Les profils électrophorétiques de tous les échantillons exposés dans les mêmes conditions sont identiques. Nous avons ainsi choisi de ne montrer qu'un seul gel contenant 6 profils électrophorétiques par condition d'exposition. La figure 78 permet de constater l'absence de contaminants majoritaires dans l'ensemble des 45 extraits histoniques.

I.3.2 Profilage LC-MS des histones extraites

Pour constituer l'échantillon « contrôle qualité » ou QC, 3 µL de chaque échantillon préalablement dilué à une concentration de 0,3 µg/µL ont été prélevés puis mélangés. L'ordre d'injection des 45 échantillons a été orthogonalisé selon la procédure précedemment décrite. Environ 1,5 µg de chaque échantillon a été injecté, et 5 µL de l'échantillon QC ont été injectés tous les 5 échantillons.

Les profils chromatographiques de trois échantillons représentatifs de chacun des groupes sont représentés figure 85. Quelle que soit la condition d'exposition, les mêmes espèces d'histones de cœur sont présentes sur les chromatogrammes, à savoir H4, H2B, H2A1, H2A2 et H3.1. Seules les abondances relatives de certains types d'histones semblent varier.

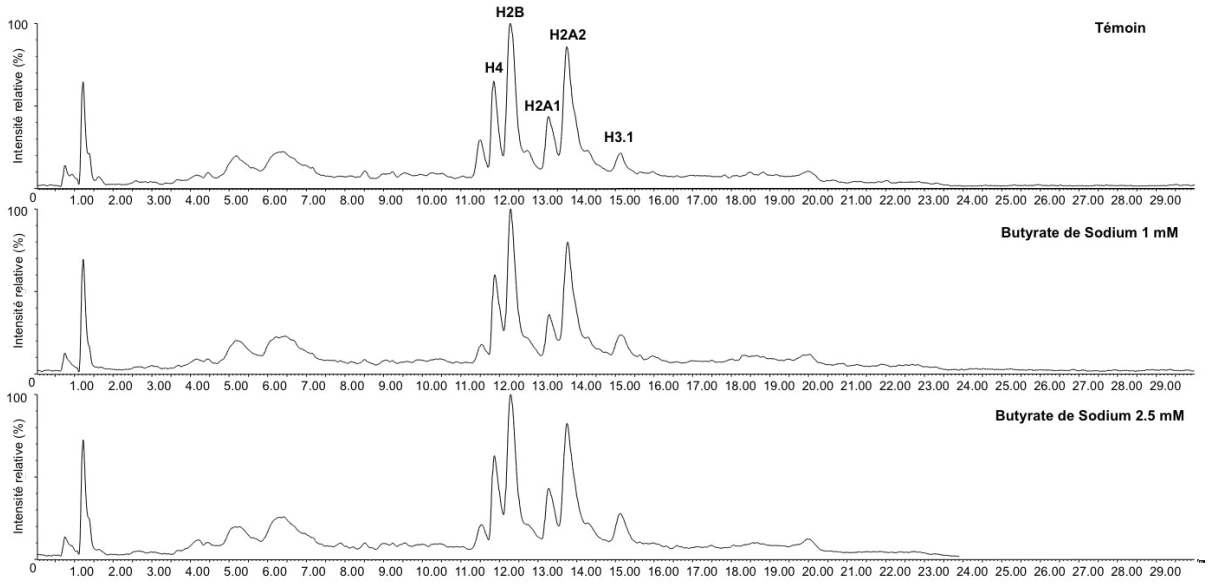


Figure 85 : profils chromatographiques obtenus en UPLC-ESI-QTOF de trois échantillons représentatifs de chacun des groupes : témoin, BS 1 mM et BS 2,5 mM.

D'après les chromatogrammes, toutes les histones sont éluées entre 10 et 16 minutes, ce qui nous permet de nous focaliser sur cette fenêtre de temps de rétention pour la suite de l'analyse. Il est possible de reconstituer un spectre MS pour chaque pic chromatographique afin d'examiner les séries d'ions multichargés des différentes espèces présentes sous ce pic.

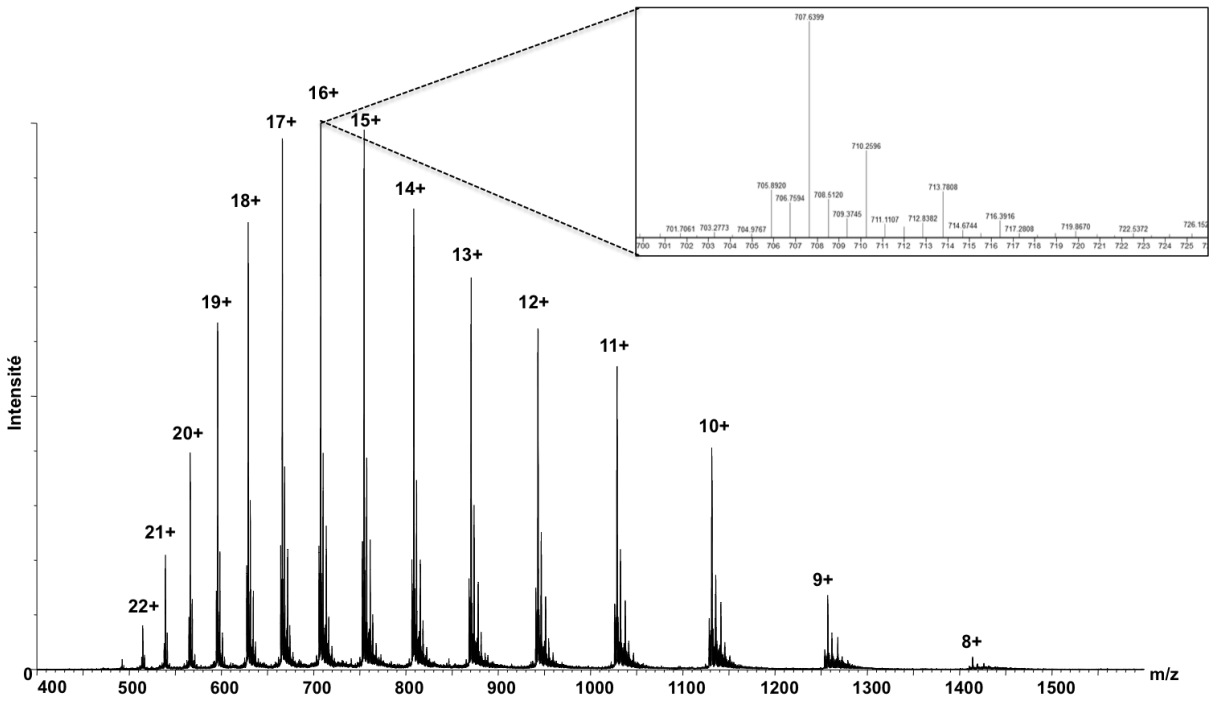


Figure 86 : exemple d'un spectre MS centroïde de l'histone H4 extraite d'un échantillon témoin. La distribution des états de charge s'étale de +8 à +22.

I.4 Prétraitement et normalisation des données

I.4.1 Prétraitement par XCMS

Une fois l'ensemble des fichiers bruts convertis au format mzData, ils ont été prétraités sous R à l'aide de XCMS. La déviation des temps de rétention entre les différents échantillons était très faible tout au long de l'analyse, comme l'atteste la figure 87 qui présente la déviation en fonction du temps de rétention.

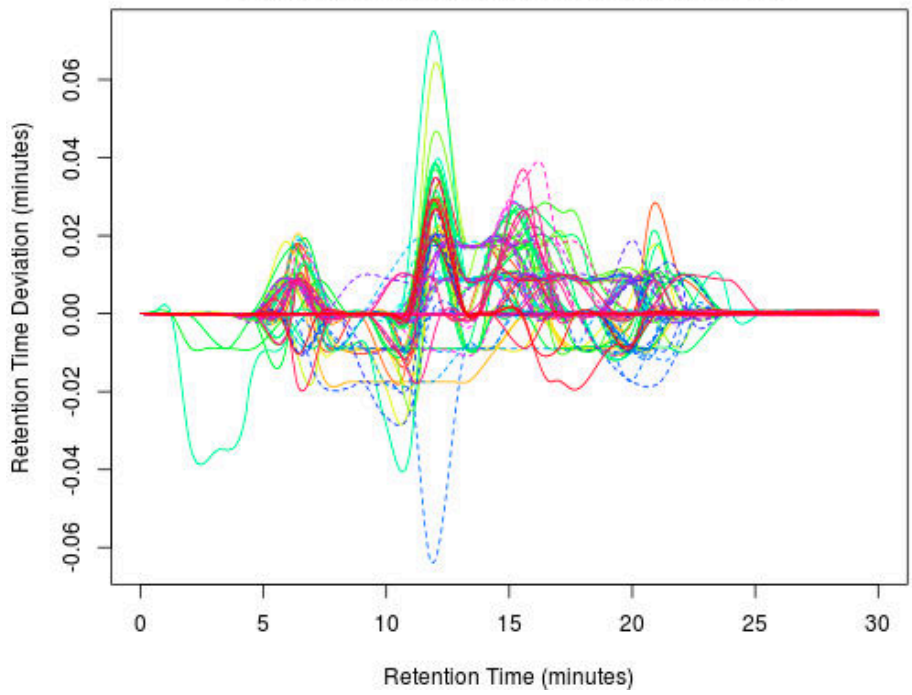


Figure 87 : déviation du temps de rétention observée en fonction du temps de rétention pour l'ensemble des 45 échantillons prétraités. Chaque ligne de couleur représente un échantillon.

Les différents chromatogrammes ont ensuite été réalignés en utilisant l'algorithme obiwarp intégré à XCMS (figure 88).

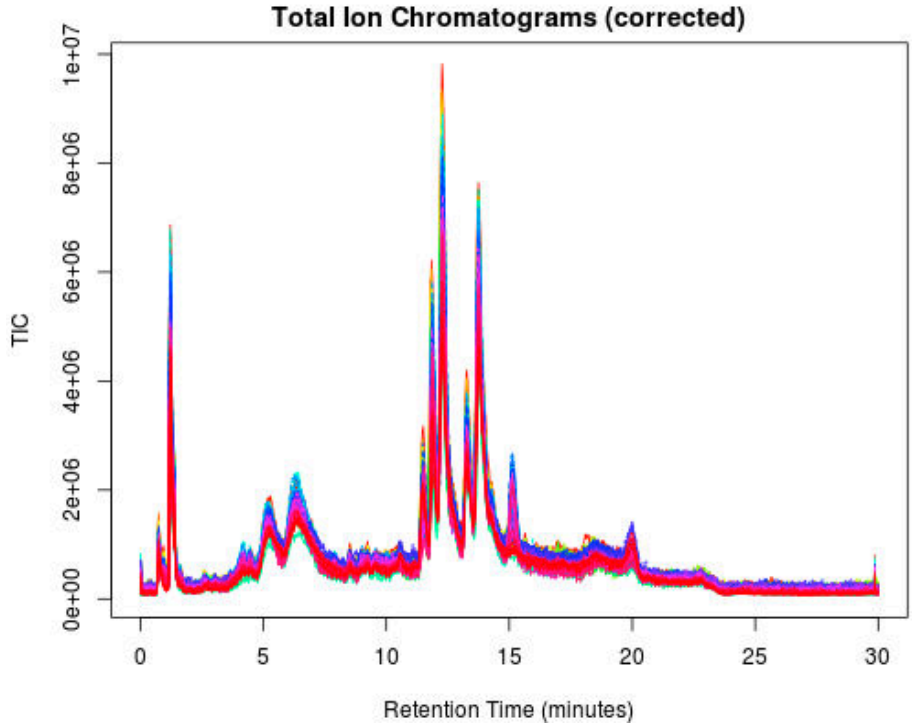


Figure 88 : superposition de l'ensemble des chromatogrammes (TIC) après réalignement et correction des temps de rétention. Chaque ligne de couleur représente le chromatogramme d'un échantillon.

Au final, la matrice *X* obtenue présente 10 004 variables différentes identifiées par leur couple t_R_*m/z*. Comme indiqué précédemment, l'ensemble des histones de cœur sont éluées entre 10 et 16 minutes. Toutes les variables dont le temps de rétention était en dehors de cette plage ont été retirées de la matrice *X*. Cette première étape nous a permis de réduire la matrice à 8 537 variables.

A partir de cette matrice affinée, nous avons voulu évaluer le niveau global de précision de notre méthode en calculant les CV (%) pour chaque variable détectée à travers tous les réplicats QC, puis en prenant la médiane de ces valeurs. Nous avons ainsi obtenu une valeur médiane de 18,8%, donc inférieure à 20% qui est la valeur limite préconisée par la FDA dans le cadre d'une étude de biomarqueurs[254]. Le niveau de précision global de notre méthode dans ce cas est donc jugé acceptable.

I.4.2 Normalisation des données

La stratégie de normalisation des données visait à rendre les échantillons comparables entre eux et à supprimer toutes les sources de variabilité non-induite pour se concentrer sur la variabilité biologique induite par l'exposition au butyrate de sodium. La normalisation par la médiane, la transformation logarithmique et le redimensionnement de Pareto ont donc été appliqués à l'ensemble des échantillons contenus dans la matrice *X*. L'évaluation des étapes de normalisation se fait à l'aide de différents types de représentations graphiques, le but étant d'obtenir une distribution gaussienne de l'intensité des variables à travers les échantillons. La figure 89 représente sous forme de boîtes à moustaches ou *box plots* les caractéristiques de position de l'intensité de plusieurs variables sélectionnées aléatoirement parmi les 8 537 contenues dans la matrice *X*. Avant normalisation, les caractéristiques de ces variables sont très hétérogènes, et certaines variables de faible intensité semblent écrasées par celles de forte intensité. Après les étapes de normalisation, ces caractéristiques sont beaucoup plus homogènes et comparables entre elles.

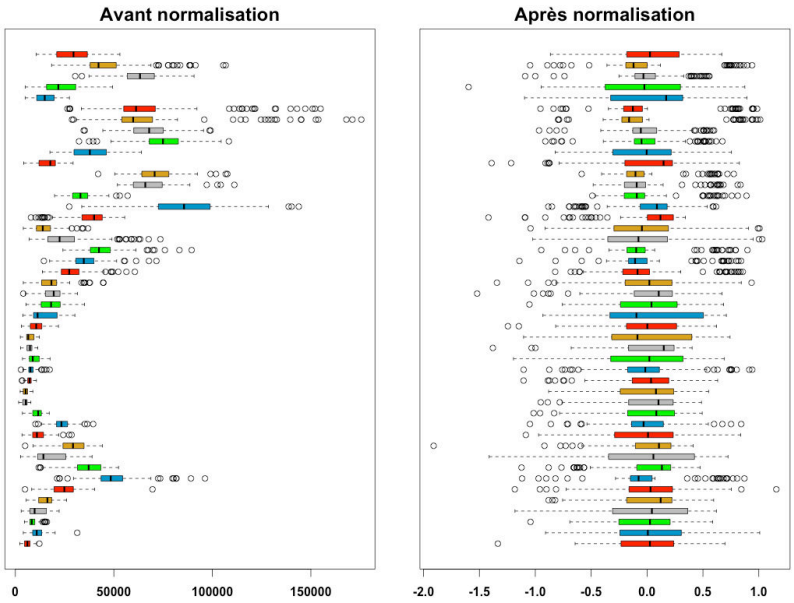


Figure 89 : boîtes à moustaches ou *box plots* résumant les caractéristiques avant et après normalisation de 50 variables sélectionnées aléatoirement parmi les 8 537 variables de la matrice *X*. L'intensité des variables est représentée en abscisse.

La densité de probabilité de l'intensité des 8 537 variables a également été évaluée pour s'assurer qu'elle était distribuée selon une gaussienne. Ainsi, la figure 90 montre l'influence de la normalisation de cette distribution. Avant normalisation, la distribution était très loin d'une gaussienne et l'amplitude des intensités est très importante. Après normalisation, la distribution est gaussienne et centrée autour de 0.

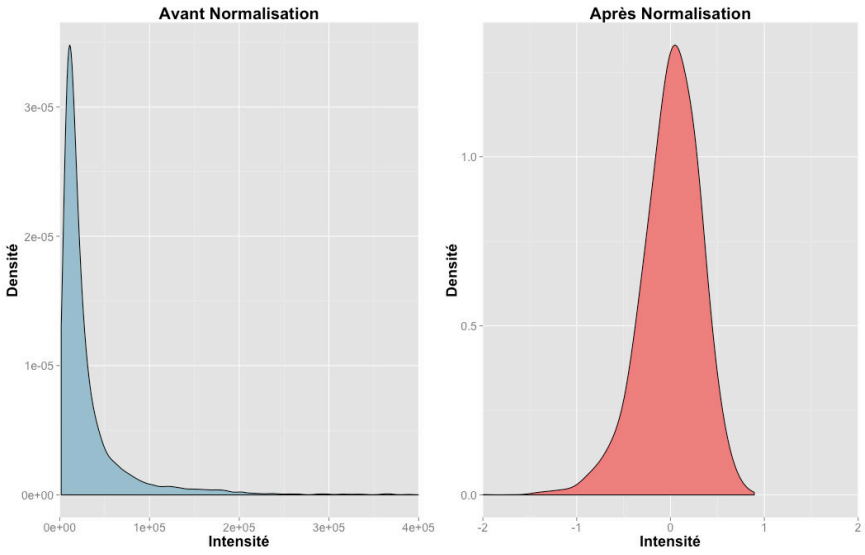


Figure 90 : estimation par noyau de la densité de probabilité de l'intensité des 8 537 variables de la matrice *X* avant (gauche) et après (droite) normalisation.

Après normalisation les données sont donc prêtes à être explorées par différentes méthodes d'analyse statistique multivariée.

I.5 Analyses statistiques descriptives

Avant de débuter l'analyse des données, la matrice X a été scindée en deux jeux de données : un jeu de données d'apprentissage et un jeu de données de prédiction. Pour cela, 5 échantillons de chacune des trois classes ont été aléatoirement retirés de la matrice initiale pour constituer le jeu de données de prédiction. Le jeu de données d'apprentissage sur lequel l'ensemble des analyses statistiques a été effectué est donc composé de 10 échantillons par classe soit 30 échantillons au total, sans compter les réplicats QC.

I.5.1 Classification ascendante hiérarchique

La classification ascendante hiérarchique a été la première méthode non supervisée utilisée pour explorer la variabilité présente dans nos données. Appliquée au jeu de données d'apprentissage, elle nous a permis d'agglomérer les échantillons en fonction de la similarité de leur profil d'intensités des variables. La représentation graphique sous forme de *heat map* présentée figure 91 permet de

visualiser les 250 variables les plus significatives entre les différentes classes d'après le résultat d'un test ANOVA.

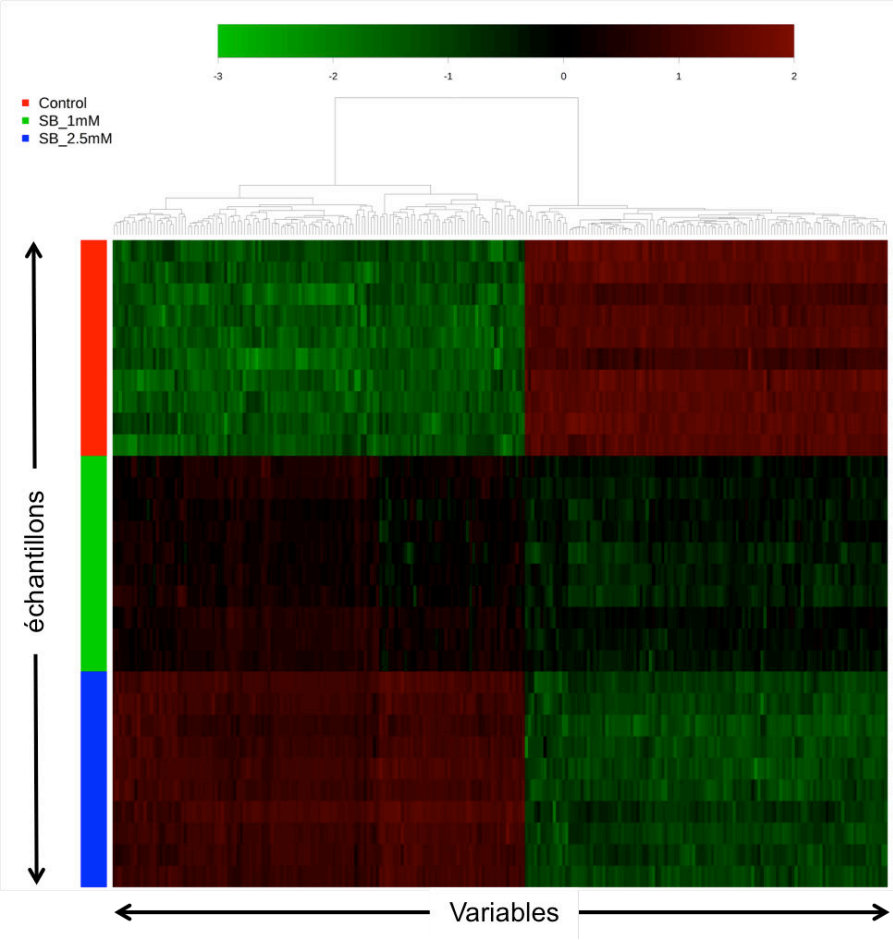


Figure 91 : classification hiérarchique ascendante et représentation *heat map* des 250 variables les plus significatives entre les 3 classes d'après un test ANOVA. Les 10 réplicats biologiques pour chaque classe d'échantillons constituant le jeu de données d'apprentissage sont représentés sur l'axe vertical. Chaque classe d'échantillons est identifiée par une couleur : rouge = témoins, vert = BS 1 mM et bleu = BS 2,5 mM. Le gradient de couleur a été utilisé pour représenter les intensités normalisées des variables sur une échelle logarithmique. La couleur verte indique une sous-expression des variables tandis que la couleur rouge indique une surexpression des variables.

Comme l'attestent les blocs de couleur sur l'axe vertical, les échantillons appartenant à la même classe ont naturellement été regroupés entre eux. En parallèle, les deux branches les plus longues du dendrogramme (maximum de dissimilarité) mettent nettement en évidence deux groupes de variables dont les intensités sont très différentes à travers les différentes classes d'échantillons. D'après le gradient de couleur, les différences d'intensité sont substantielles. La branche gauche du dendrogramme englobe les variables qui sont surexprimées de manière dose-dépendante par une exposition au butyrate de sodium. Celle de

droite englobe les variables qui sont sous-exprimées de manière dose-dépendante par une exposition au butyrate de sodium.

Bien que les résultats de la classification ascendante hiérarchique nous permettent d'affirmer qu'il existe une différence significative entre les profils d'histones des échantillons témoins et traités au butyrate de sodium, le type de représentation graphique associé ne facilite pas l'extraction des variables discriminantes. La figure 91 ne présente que 250 des 8 537 variables contenues dans la matrice X, car il serait impossible de visualiser sur un même graphique l'intégralité de ces variables simultanément. Elle représente donc une première approche intéressante mais trouve ses limites face au nombre très important de variables.

I.5.2 Analyse en composantes principales

L'analyse en composantes principales (ACP) nous permet d'explorer plus en profondeur la variabilité naturelle qui existe entre les profils d'histones des échantillons appartenant aux différentes classes. En réduisant la dimensionnalité de la matrice X, elle résume la majorité de la variabilité sur quelques composantes principales. Le nombre de composantes principales significatives a été fixé à trois d'après la procédure de validation croisée. Ces trois composantes permettent de capter 35,5 % de la variance totale (R^2X_{cum} = 0,355). La figure 92 représente la répartition des échantillons dans un plan en 3D défini par les 3 composantes principales. Ce *scores plot* 3D permet de visualiser la répartition naturelle des échantillons dans l'espace et de repérer les tendances naturelles de regroupement entre les échantillons.

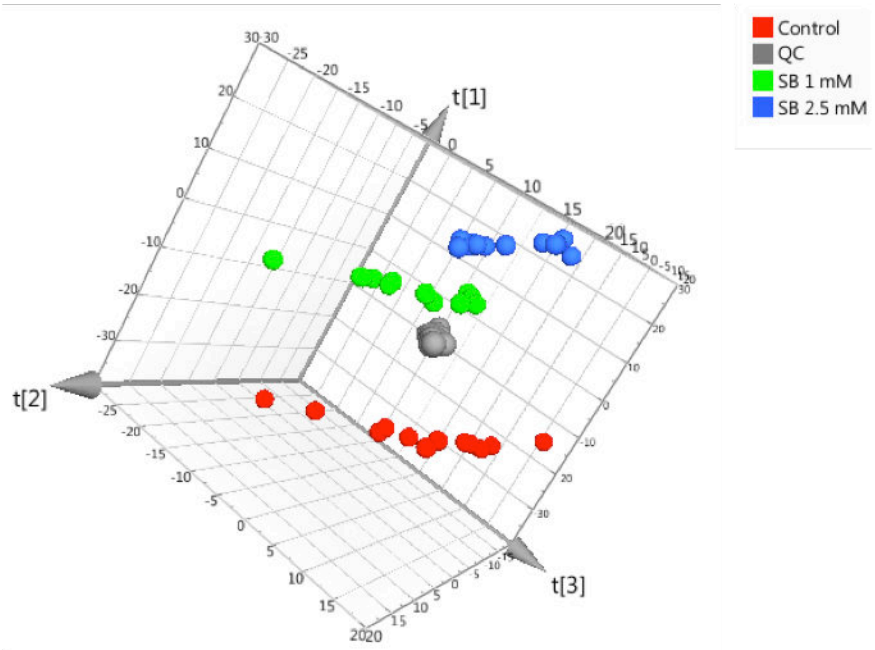


Figure 92 : *scores plot* 3D d'une ACP représentant les 3 composantes sélectionnées. Les quatre classes d'échantillons sont représentées par des couleurs différentes : rouge = témoins, vert = BS 1 mM, rouge = BS 2,5 mM et gris = QC.

En utilisant la représentation graphique appelée DModX (*Distance to the Model in X space*) proposée par le logiciel SIMCA-13 (figure 93) nous pouvons évaluer la distance de chacun des individus par rapport à la première composante principale du modèle ACP. Nous pouvons ainsi constater qu'aucun individu aberrant n'est présent dans notre série d'échantillons et que tous ont un comportement homogène.

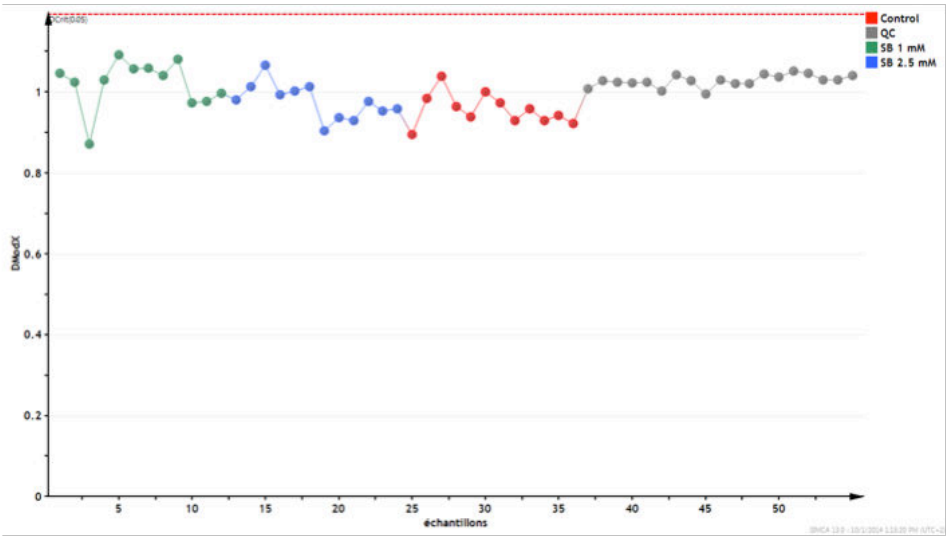


Figure 93 : représentation DModX de la distance des individus par rapport à la première composante principale du modèle ACP. Chaque point de couleur représente un échantillon. Le code couleur est le même que précédemment.

Pour simplifier l'interprétation du modèle ACP à trois composantes, le *scores plot* 2D qui est présenté figure 94 ne conserve que les deux premières composantes principales PC1 et PC2, la troisième ne résumant que peu de variabilité.

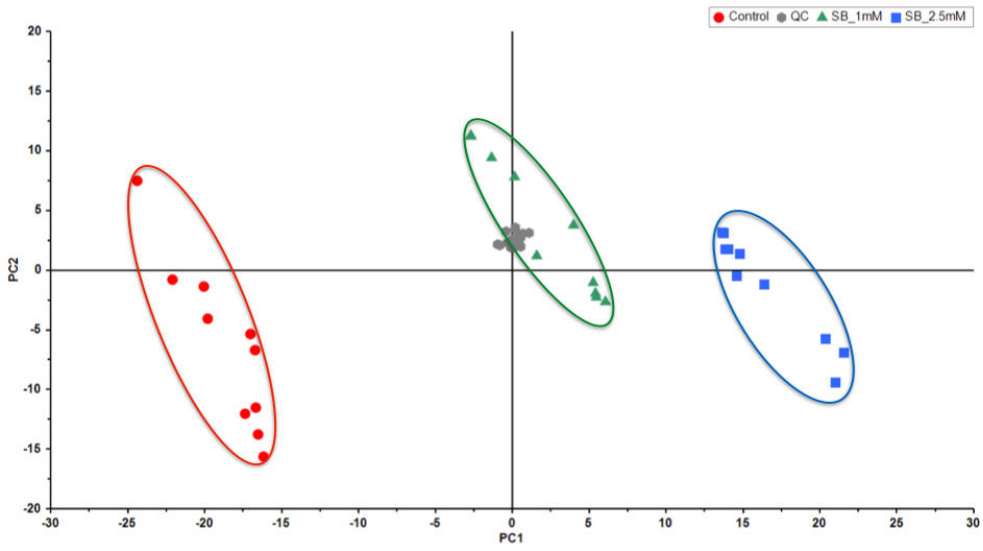


Figure 94 : *scores plot* 2D d'une ACP représentant les deux premières composantes. Les quatre classes d'échantillons sont représentées par des couleurs et des formes différentes : ○ = témoins, △ = BS 1 mM, □= BS 2,5 mM et ⬢ = QC.

À partir de la figure 94, nous pouvons constater que les réplicats de l'échantillon QC sont parfaitement groupés au centre du plan, ce qui atteste de la

qualité et de la stabilité de l'analyse. Nous pouvons également observer un regroupement net des échantillons en fonction de leur condition d'exposition au butyrate de sodium. La première composante principale (PC1) permet de séparer les trois groupes d'échantillons formés naturellement, et prouve que la principale source de variabilité dans le jeu de données correspond aux conditions d'exposition. De plus, la PC1 souligne un effet dose-réponse en séparant clairement les échantillons exposés au butyrate de sodium à 1 mM de ceux exposés à 2,5 mM. La PC2 représente, elle, la variabilité qui existe entre les individus d'une même classe mais reste négligeable par rapport à la variabilité inter-classes.

Pour se faire une première idée générale du nombre et de la nature des variables responsables de cette séparation naturelle entre les échantillons, nous avons utilisé le *loadings plot* présenté figure 95. Cette représentation graphique utilise le même plan défini par l'ACP mais y projette non plus les individus mais les variables. Nous retrouvons ainsi les mêmes composantes et nous pouvons en extraire les variables responsables de la dispersion des échantillons sur les deux premières composantes.

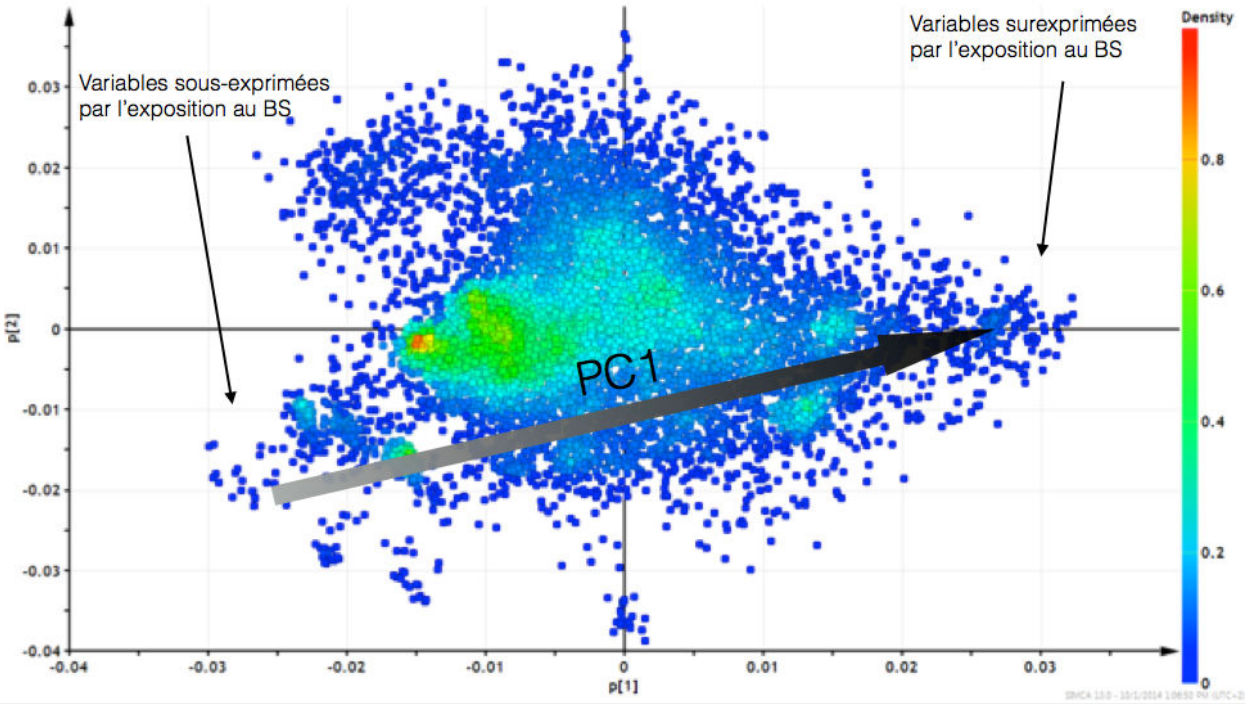


Figure 95 : l*oadings plot* du modèle ACP. Le gradient de couleur représente la densité des variables lorsqu'elles se superposent.

Nous pouvons observer sur la figure 95 un vaste nuage de points relativement dense. La majorité des 8 537 variables qui constituent ce nuage de points ne sont pas directement responsables de la séparation entre les échantillons témoins et ceux exposés au butyrate de sodium. En superposant ce *loadings plot* avec le *scores plot* de la figure 94, nous pouvons repérer les variables dont l'abondance est corrélée à la dose de butyrate de sodium. Ainsi, les variables situées à l'extrémité gauche du nuage de points sont sous-exprimées lors d'une exposition au butyrate de sodium alors que celles situées à l'extrémité droite sont surexprimées de manière dose-dépendante. D'après leur temps de rétention, la plupart des variables surexprimées semblent correspondre à des formes différentiellement modifiées des histones H4 et H2B. Cependant, vu le très grand nombre de variables il est difficile de les hiérarchiser et de déterminer leur importance réelle dans la séparation des groupes en utilisant seulement le *loadings plot*. De plus, les points restent relativement groupés et aucun marqueur réel ne semble émerger.

I.5.3 Conclusion

Les résultats des analyses descriptives non supervisées révèlent qu'une exposition des cellules BeWo au butyrate de sodium induit un changement significatif de leur profil d'histones, et que ce changement est dose-dépendant. Ils permettent également de remarquer que certains sous-types d'histones semblent davantage impactés que d'autres par l'exposition au butyrate de sodium, notamment H4 et H2B. Les modèles statistiques non supervisés ne sont cependant pas suffisants pour extraire les variables les plus discriminantes entre les groupes d'échantillons.

I.6 Classification des échantillons et analyses statistiques prédictives

En complément des analyses non supervisées, la construction de modèles prédictifs robustes poursuit deux objectifs : la classification des échantillons et l'extraction des variables discriminantes. Ces modèles pourront ensuite servir à prédire l'appartenance de nouveaux échantillons à une des classes en fonction des similitudes observées.

I.6.1 Analyse supervisée PLS-DA des trois classes

L'analyse discriminante PLS-DA nous a permis de modéliser la relation entre les profils d'histones et l'appartenance des échantillons à une des trois classes. Les échantillons contenus dans le jeu de données d'apprentissage ont été étiquetés en fonction de leur condition d'exposition. Trois classes ont ainsi été définies : la classe 1 correspond aux échantillons témoins, la classe 2 aux échantillons exposés au butyrate de sodium à 1 mM et la classe 3 à ceux exposés au butyrate de sodium à 2,5 mM. Après validation croisée, le modèle PLS-DA retenu comporte 3 composantes. Les paramètres R^2Y et Q^2Y étaient respectivement égaux à 0,99 et 0,94, attestant de la fiabilité du modèle. Le modèle a ensuite été validé selon les différentes procédures détaillées en partie expérimentale. Le test de permutation a été effectué sur 999 itérations et a aboutit à une droite de régression de Q^2Y coupant l'axe des ordonnées en dessous de 0 (figure 96). De plus, toutes les valeurs de R^2Y et Q^2Y obtenues pour les modèles générés par permutation sont inférieures à celles du modèle original.

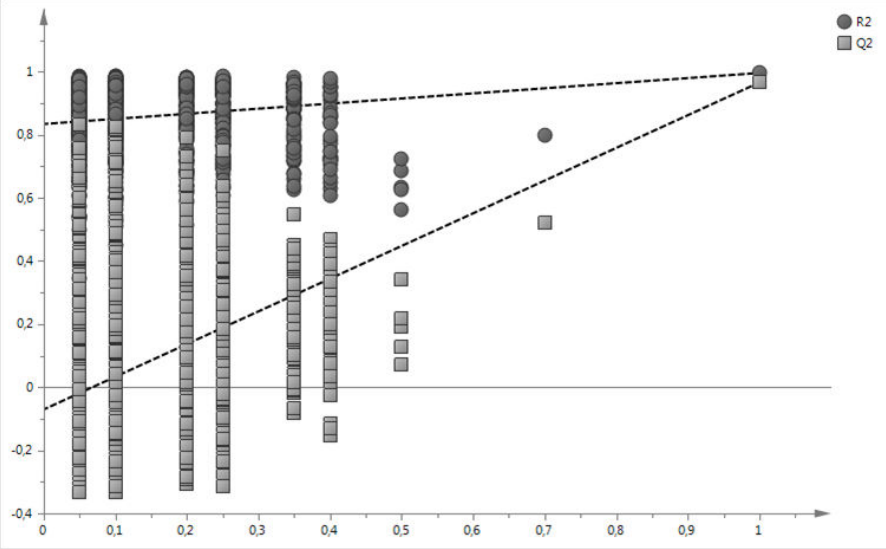


Figure 96 : résultat du test de permutation présentant les droites de régression de R^2 et Q^2 des modèles générés aléatoirement.

La *p-value* obtenue par le test CV-ANOVA est égale à $1,8 \times 10^{-14}$, c'est-à-dire bien en dessous du seuil d'acceptabilité de 0,001. Le modèle a donc été parfaitement validé par les méthodes internes. À partir de là, le *scores plot* du modèle PLS-DA présenté figure 97 a été interprété. Il montre la nette séparation

entres les trois classes. Deux composantes sont responsables de la discrimination des classes : la première est responsable de la discrimination en fonction de l'exposition au butyrate de sodium, et la deuxième en fonction de la dose de butyrate de sodium.

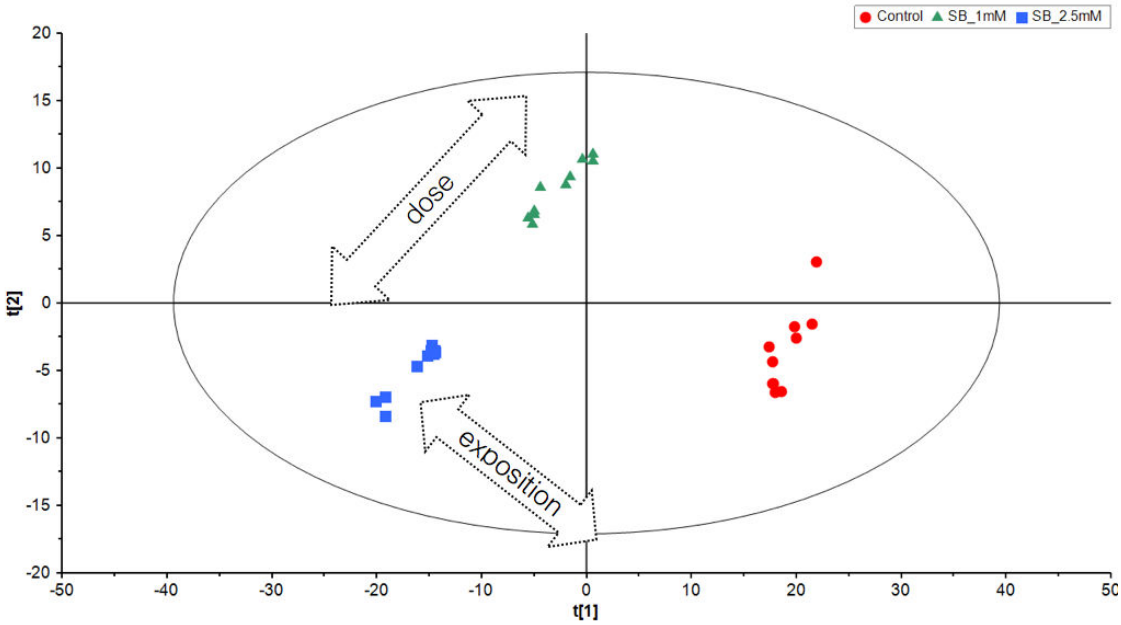


Figure 97 : *scores plot* du modèle PLS-DA obtenu à partir du jeu de données d'apprentissage.

Nous avons ensuite testé la capacité du modèle PLS-DA à classer des nouveaux échantillons inconnus non étiquetés. Nous avons utilisé pour cela le jeu de données de prédiction composé de cinq échantillons par classe. En projetant ces nouveaux échantillons dans l'espace défini à l'aide du jeu de données d'apprentissage, nous avons observé leur répartition. La figure 98 présente les résultats graphiques de cette projection et permet d'observer la discrimination des échantillons en fonction de l'exposition. Cependant la discrimination en fonction de la dose ne semble pas être conservée. En effet, les échantillons exposés au butyrate de sodium à 1 mM et à 2,5 mM ne sont pas séparés selon la composante « dose ».

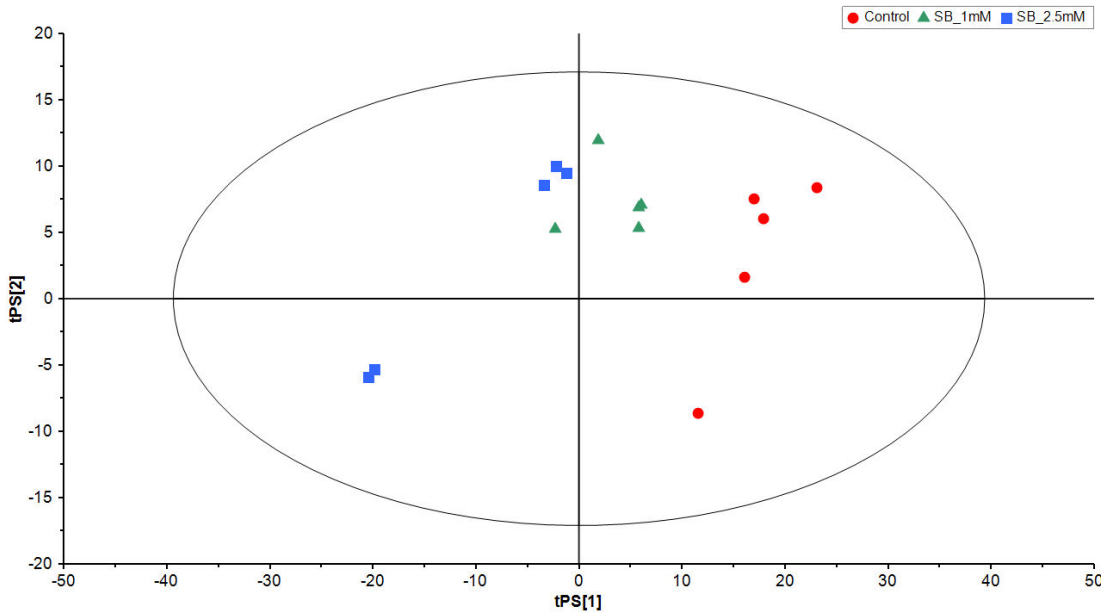


Figure 98 : *scores plot* de la projection dans le modèle PLS-DA défini précédemment du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 ou 2,5 mM.

Pour vérifier cela avec des chiffres, nous avons utilisé la table des erreurs de classification générée avec la prédiction (tableau 21). Cette table confirme les hypothèses émises à la vue du *scores plot*.

Tableau 21 : erreurs observées après classification par le modèle PLS-DA du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 ou 2,5 mM.

	Membres	Corrects	Témoin	BS 1 mM	BS 2,5 mM	Pas de classe
Témoin	5	100%	5	0	0	0
BS 1 mM	5	100%	0	5	0	0
BS 2,5 mM	5	40%	0	3	2	0
Pas de classe	0	-	0	0	0	0
Total	15	80%	5	8	2	0
Prob. Fisher	0,00044					

Parmi les cinq échantillons exposés au butyrate de sodium à 2,5 mM, trois ont été classés par le modèle comme ayant été exposés au butyrate de sodium à 1 mM. Ceci correspond à une précision de classification de seulement 40% pour cette classe. L'ensemble des échantillons pour les deux autres classes a été correctement classé. La précision finale de ce modèle a été estimée à 80%. Ce modèle est donc suffisamment performant pour discriminer les échantillons témoins et les échantillons exposés, mais ne parvient pas à classer systématiquement les échantillons en fonction de la dose de butyrate de sodium à laquelle ils ont été exposés.

I.6.2 Analyses supervisées binaires OPLS-DA

I.6.2.1 Témoin versus butyrate de sodium 1 mM

Face aux limites rencontrées dans le cas du modèle PLS-DA à trois classes, nous avons choisi de construire un modèle OPLS-DA pour chacune des deux doses de butyrate de sodium, puis de comparer les variables discriminantes qui en seront extraites. Le premier modèle OPLS-DA a été construit à partir d'un jeu de données d'apprentissage contenant les 10 échantillons témoins et les 10 échantillons exposés au butyrate de sodium à 1 mM. Le modèle généré comporte une composante prédictive et une composante orthogonale. Les paramètres statistiques du modèle sont les suivants : $R^2Y = 0,99$, $Q^2Y = 0,97$. Le test CV-ANOVA nous a fourni une *p-value* de $4,5 \times 10^{-11}$. Le modèle a donc largement été validé. Le *scores plot* correspondant est présenté figure 99.

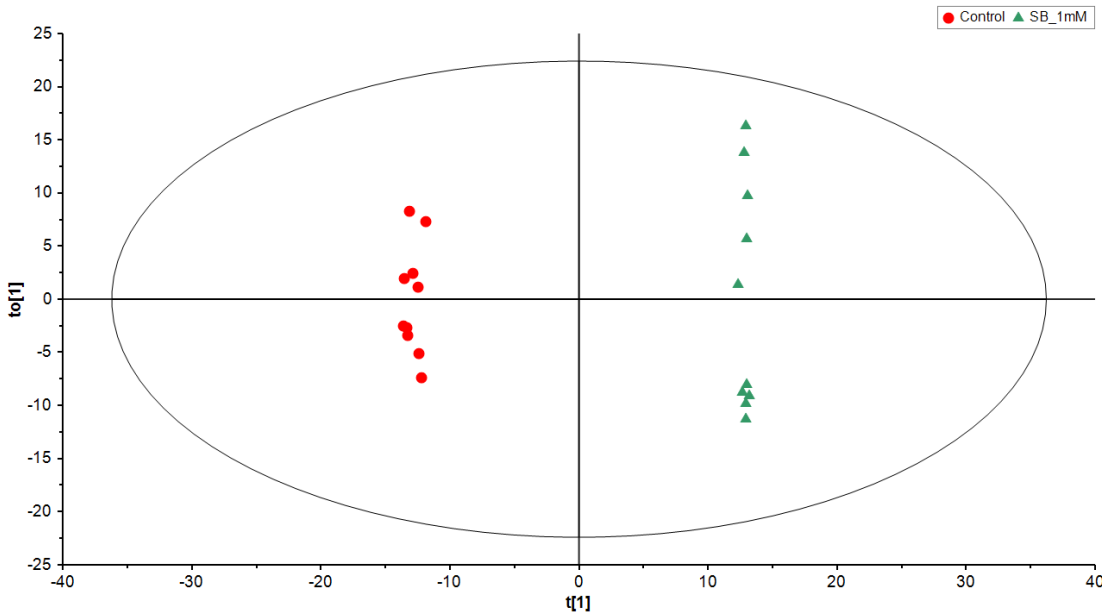


Figure 99 : *scores plot* du modèle OPLS-DA obtenu à partir du jeu de données d'apprentissage contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 mM.

Les échantillons sont donc clairement discriminés en fonction de leur classe sur la composante prédictive t[1]. La classification du jeu de données de prédiction en utilisant ce modèle a été réalisée et le résultat est présenté figure 100. La composante prédictive semble être capable de discriminer ces échantillons de manière naïve.

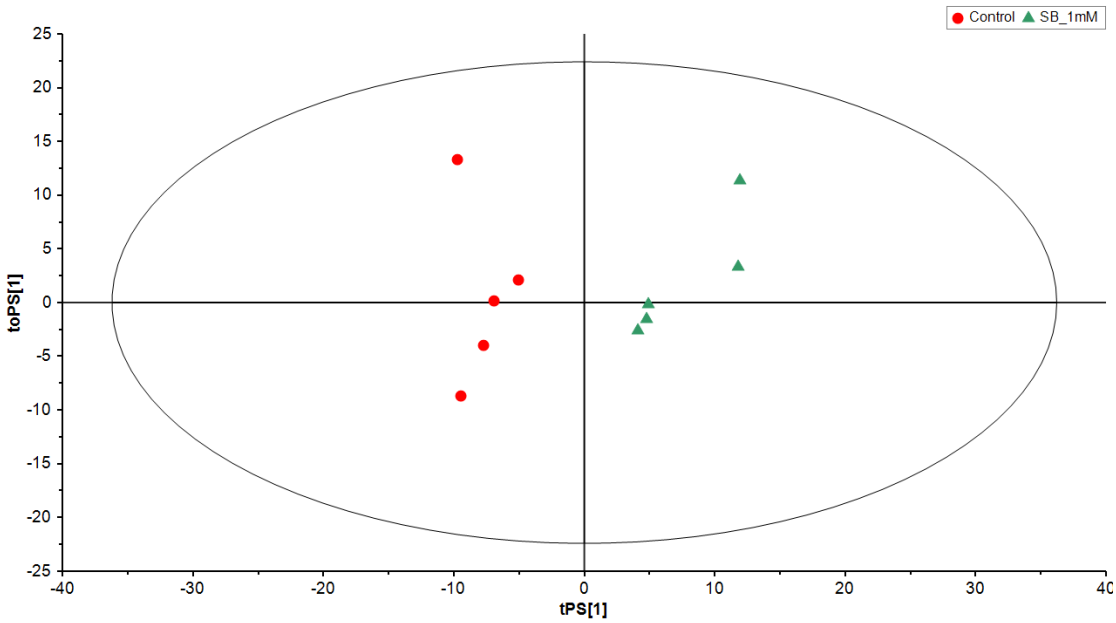


Figure 100 : *scores plot* de la projection dans le modèle OPLS-DA défini précédemment du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 mM.

Ce résultat a été confirmé à l'aide de la table des erreurs de classification présentée ci-dessous (tableau 22).

Tableau 22 : erreurs observées après classification par le modèle OPLS-DA du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 mM.

	Membres	Corrects	Témoin	BS 1 mM
Témoin	5	100%	5	0
BS 1 mM	5	100%	0	5
Pas de classe	0	-	0	0
Total	10	100%	5	5
Prob. Fisher	0,004			

L'ensemble des échantillons inconnus a été correctement classé, conduisant à une précision totale de 100%. Ce modèle est donc parfaitement valide et sera utilisé par la suite pour extraire les variables discriminantes entre ces deux classes.

I.6.2.2 Témoin versus butyrate de sodium 2,5 mM

Le second modèle OPLS-DA a été construit à partir d'un jeu de données d'apprentissage contenant les 10 échantillons témoins et les 10 échantillons exposés au butyrate de sodium à 2,5 mM. Le modèle généré présente une composante prédictive ainsi qu'une unique composante orthogonale. Les valeurs des paramètres statistiques du modèle (R^2Y = 0,99 et Q^2Y = 0,98) ainsi que la *p-value* de 9,3 x 10^{-13} fournie par le test ANOVA témoignent de la validité du modèle. Le *scores plot* correspondant est présenté figure 101.

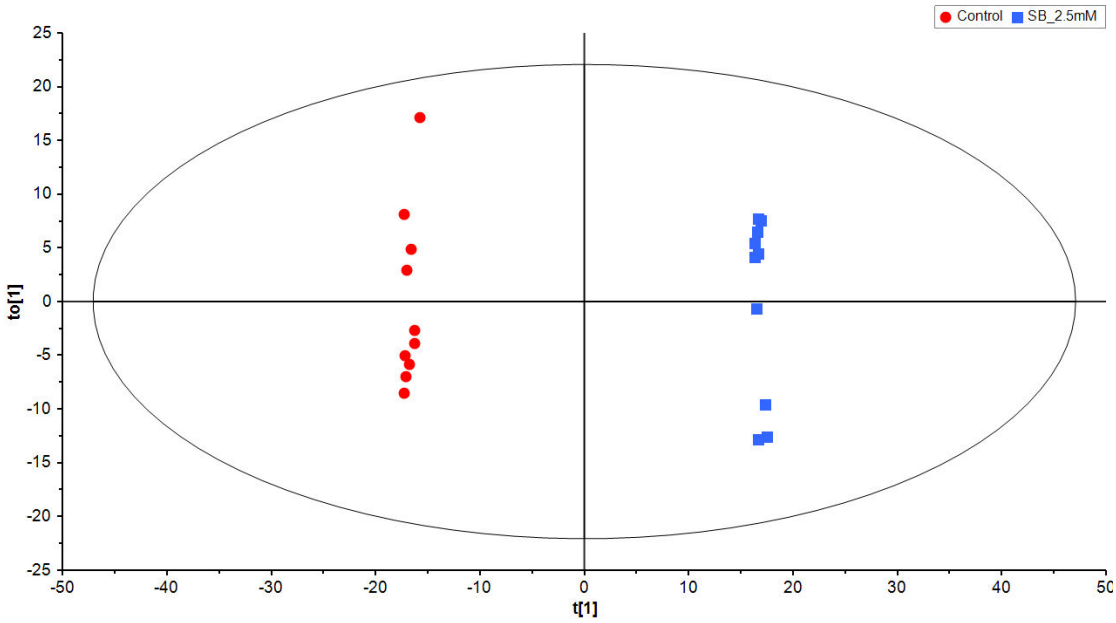


Figure 101 : *scores plot* du modèle OPLS-DA obtenu à partir du jeu de données d'apprentissage contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 2,5 mM.

Le modèle généré pour cette dose semble également capable de discriminer les échantillons en fonction de leur classe sur la composante prédictive t[1]. La classification du jeu de données de prédiction en utilisant ce modèle a été réalisée et le résultat est présenté figure 102. Là aussi, la composante prédictive semble être capable de discriminer ces échantillons de manière naïve.

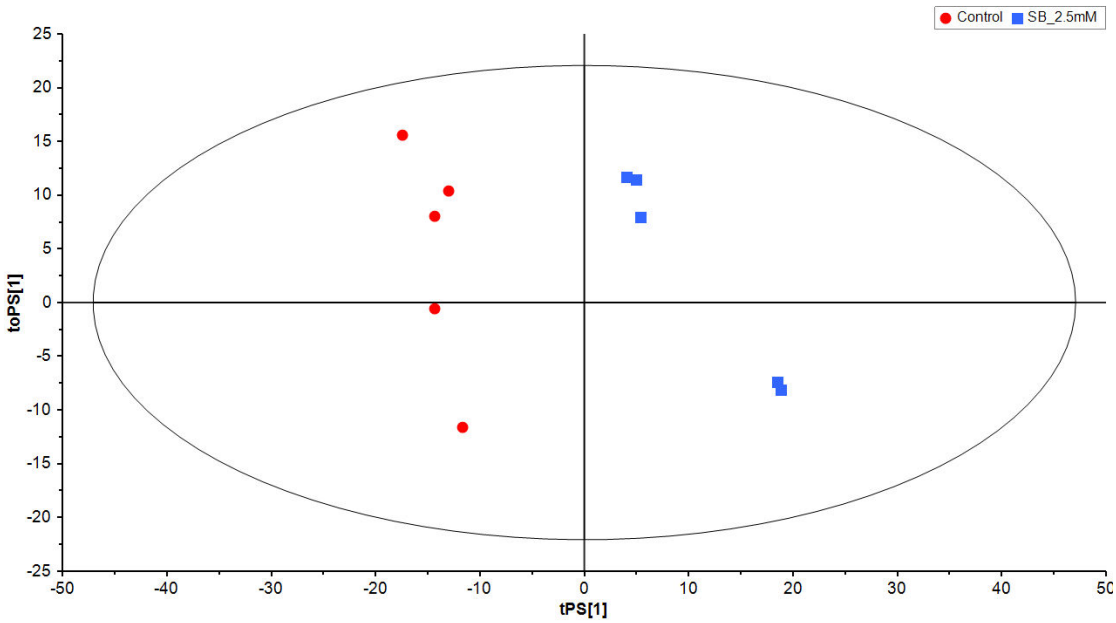


Figure 102 : *scores plot* de la projection dans le modèle OPLS-DA défini précédemment du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 2,5 mM.

En examinant la table des erreurs de classification, nous observons que l'ensemble des échantillons inconnus a été correctement classé par le modèle (tableau 23).

Tableau 23 : erreurs observées après classification par le modèle OPLS-DA du jeu de données de prédiction contenant les échantillons témoins et les échantillons exposés au butyrate de sodium à 2,5 mM.

	Membres	Corrects	Témoin	BS 2,5 mM
Témoin	5	100%	5	0
BS 2,5 mM	5	100%	0	5
Pas de classe	0	-	0	0
Total	10	100%	5	5
Prob. Fisher	0,004			

Ce modèle validé atteint une précision de classification de 100% sur le jeu de données et a donc été utilisé pour extraire les variables discriminantes entre les deux classes d'échantillons.

I.7 Formes d'histones discriminantes associées à l'exposition au butyrate de sodium

À partir des deux modèles OPLS-DA validés, nous avons extrait pour chaque dose les variables les plus discriminantes parmi les 8 537 initialement présentes. Pour cela, nous avons utilisé leur score VIP avec une valeur seuil définie à 1,5. Dans ces conditions, 155 variables ont été extraites du modèle OPLS-DA témoin *versus* butyrate de sodium 1 mM, et 115 du modèle témoin *versus* butyrate de sodium 2,5 mM. En plus du seuil de score VIP, tous les *loadings* des variables retenues avaient un coefficient de corrélation absolu (*p(corr)*) entre le modèle et les données originales supérieur à 0,6 là ou la valeur seuil habituellement utilisée est de 0,5[274].

Il existe une certaine redondance parmi les variables sélectionnées, chacune étant présente à plusieurs états de charge (jusqu'à 11 dans le cas de H4). Ainsi nous avons raffiné les deux listes de variables en ne retenant que l'état de charge le plus intense pour chaque protéoforme. Nous avons ensuite utilisé le temps de rétention et le rapport *m/z* de chacune de ces variables pour les identifier sur les spectres continuum. Après déconvolution, nous avons pu assigner une masse moléculaire moyenne à chacune d'entre elles. Un exemple de spectre déconvolué pour chaque sous-type d'histones de cœur identifié est présenté figure 103.

Ensuite, le moteur de recherche TagIdent a été utilisé pour comparer les masses moléculaires moyennes mesurées avec les masses moléculaires moyennes théoriques présentes dans la base de données de séquences protéiques UniProtKB en utilisant les paramètres décrits en partie expérimentale. Pour chaque recherche effectuée, seules des histones ont été proposées par TagIdent (figure 104) avec un maximum de deux formes différentes dans certains cas. Une fois la forme non modifiée post-traductionnellement identifiée, les formes modifiées ont été déduites en calculant les incréments de masse observés. Chaque forme modifiée ainsi proposée est référencée sur la base de données HIstome et sur la banque de

séquences protéiques UniProtKB. Les résultats pour l'exposition au BS à 1 mM sont présentés tableau 24 et ceux pour la dose à 2,5 mM sont présentés tableau 25.

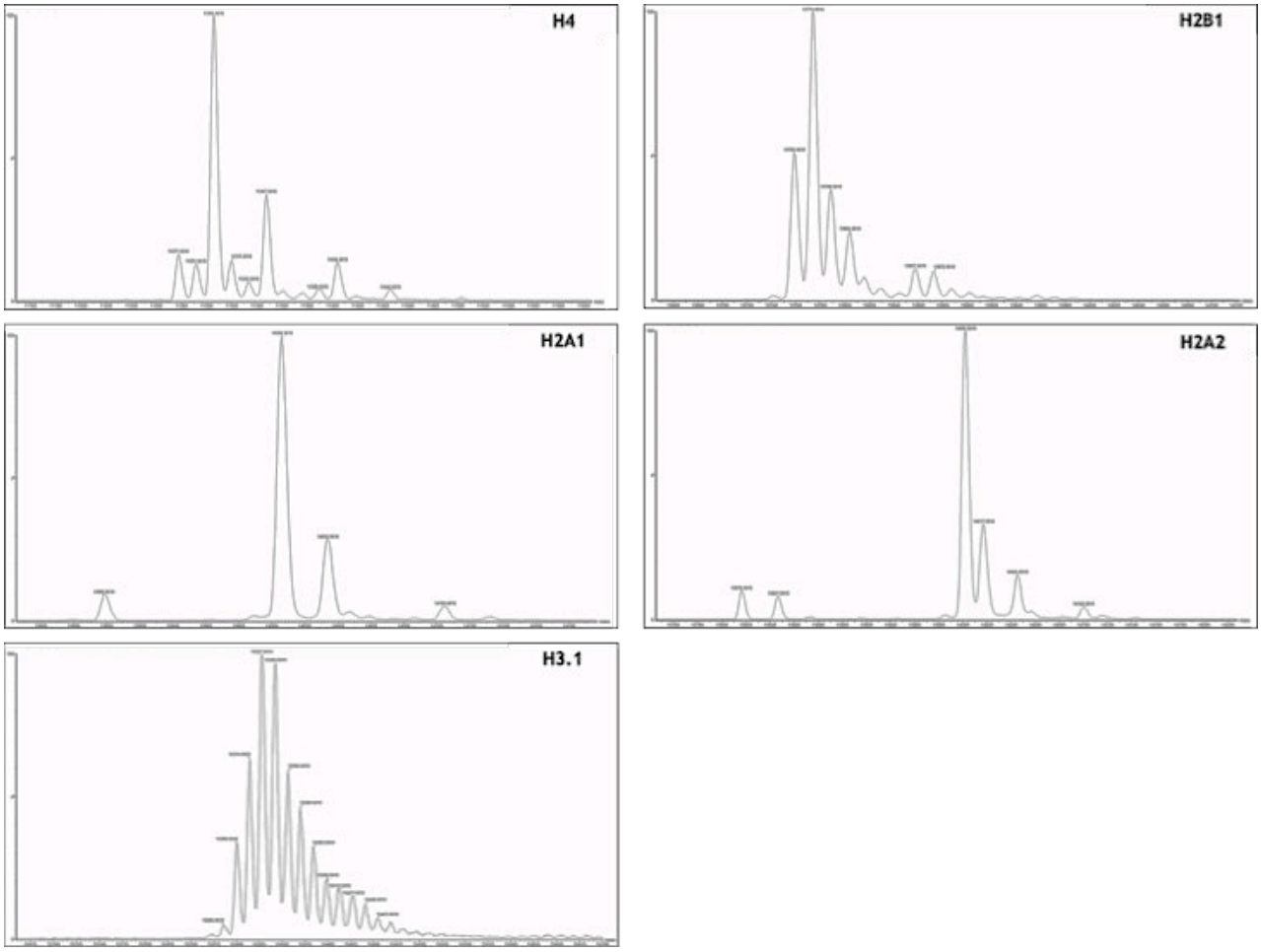


Figure 103 : spectres de déconvolution des différents sous-types d'histones de cœur identifiés.

ExPASy Bioinformatics Resource Portal — TagIdent

TagIdent

Results from TagIdent

The search in UniProtKB/Swiss-Prot has been launched with the following values:
pI range: 9 - 12
Mw range: 13757.812075 - 13759.187925
with cysteines in reduced form
OS/OC/OX = Homo Sapiens (corresponding to TaxID: 9606)
KW keyword = ALL

Scan done on 21-Oct-2014.
UniProtKB/Swiss-Prot Release 2014_09 of 01-Oct-14: 546439 entries

	Swiss-Prot
Number of proteins found in the specified pI/Mw ranges	1

Scan in UniProtKB/Swiss-Prot database (546439 entries)

1 proteins found in the specified pI/Mw range

H2B1K_HUMAN (O60814)
Histone H2B type 1-K.
Chain: 2-126, pI: 10.32, Mw: 13759

Figure 104 : exemple d'une recherche TagIdent pour la protéine de masse moyenne observée égale à 13 758,5.

Tableau 24 : identifications les plus probables des variables discriminantes entre les échantillons témoins et les échantillons exposés au butyrate de sodium à 1 mM

Identité	N° d'accès UniProtKB	Masse moyenne théorique de la forme non modifiée (Da)	Masse moyenne observée de la forme non modifiée (Da)	Incrément de masse observé (Da)	PTMs	Score VIP	CV (%)	FDR (q-value)	Ratio
H2A-1B/E	P04908	14 004,3	14 005	+42	1 ac	1,62	9,3	$8,7 \times 10^{-08}$	0,46
H2B-1K	O60814	13 758,9	13 758,5	0	***	1,5	11,9	$2,0 \times 10^{-07}$	-0,62
				+42	1 ac	1,65	16,7	$6,1 \times 10^{-07}$	0,70
				+56	1 ac + 1 me1	2,07	4,9	$2,0 \times 10^{-14}$	0,73
				+70	1 ac + 2 me1 / 1 me2	2,27	9,6	$2,3 \times 10^{-09}$	0,82
				+84	2 ac	2,37	9,6	$2,9 \times 10^{-07}$	0,90
H2B-1M	Q99879	13 858,0	13 857,5	+84	1 ac + 1 me2 + 1 me1	1,79	18	$7,8 \times 10^{-05}$	0,68
H4	P62805	11 236,1	11 236,5	+42	1 ac	2,51	16,7	$3,9 \times 10^{-14}$	-1,15
				+56	1 ac + 1me1	2,13	9	$9,1 \times 10^{-13}$	-0,79
				+70	1 ac + 1me2	1,64	4,5	$6,7 \times 10^{-14}$	-0,45
				+112	2 ac + 1me2	1,77	5,2	$6,5 \times 10^{-13}$	0,57
				+126	3 ac	3,01	11,3	$7,9 \times 10^{-09}$	1,63
				+154	3 ac + 1me2	3,08	7,7	$8,1 \times 10^{-13}$	1,66
				+168	4 ac	2,84	7,6	$9,4 \times 10^{-11}$	1,30
				+196	4 ac + 1me2	3,19	16,3	$4,7 \times 10^{-14}$	1,74
				+210	5 ac	2,08	9,6	$4,8 \times 10^{-09}$	0,80
				+252	6 ac	2,81	12,7	$1,1 \times 10^{-07}$	1,37

Tableau 25 : identifications les plus probables des variables discriminantes entre les échantillons témoins et les échantillons exposés au butyrate de sodium à 2,5 mM.

Identité	N° d'accès UniProtKB	Masse moyenne théorique de la forme non modifiée (Da)	Masse moyenne observée de la forme non modifiée (Da)	Incrément de masse observé (Da)	PTMs	Score VIP	CV (%)	FDR (q-value)	Ratio
H2A-1B/E	P04908	14 004,3	14 005	+42	1 ac	1,82	14,5	$1,6 \times 10^{-10}$	0,79
				+56	1 ac + 1 me1	1,80	19,5	$2,8 \times 10^{-10}$	1,01
				+84	2 ac	1,81	27,3	$2,3 \times 10^{-05}$	0,88
H2B-1K	O6081	13 758,9	13 758,5	+56	1 ac + 1 me1	1,79	5,0	$2,5 \times 10^{-16}$	0,80
				+70	1 ac + 2 me1 / 1 me2	2,06	8,9	$2,8 \times 10^{-15}$	1,08
				+84	2 ac	2,12	8,9	$5,4 \times 10^{-11}$	1,13
H3.1	P68431	15 272,9	15 272,5	+28	1 me2	1,79	23,6	$2,6 \times 10^{-06}$	-1,22
				+42	1 ac	1,83	16,9	$1,5 \times 10^{-07}$	-1,40
				+56	1 ac + 1 me1	2,10	24,8	$5,1 \times 10^{-07}$	-1,51
				+70	1 ac + 1 me2	2,06	19,0	$2,6 \times 10^{-06}$	-1,56
				+98	2 ac + 1 me1	1,80	25,2	$8,6 \times 10^{-04}$	1,01
				+112	2 ac + 2 me1 / 1 me2	1,76	29,1	$8,9 \times 10^{-05}$	1,16
				+154	3 ac + 2 me1 / 1 me2	1,77	15,9	$7,8 \times 10^{-05}$	1,12
H4	P62805	11 236,1	11 236,5	+42	1 ac	2,05	8,5	$4,0 \times 10^{-17}$	-1,36
				+56	1 ac + 1me1	1,85	12,6	$2,0 \times 10^{-15}$	-1,15
				+126	3 ac	2,45	15,4	$3,3 \times 10^{-11}$	1,69
				+140	3 ac + 1me1	2,41	18,3	$1,8 \times 10^{-09}$	1,58
				+154	3 ac + 1 me2	2,67	11,4	$7,3 \times 10^{-16}$	1,98
				+168	4 ac	2,35	9,0	$6,5 \times 10^{-15}$	1,48
				+196	4 ac + 1 me2	2,91	12,4	$4,2 \times 10^{-18}$	2,30

L'ensemble des paramètres statistiques multivariés et univariés confirme la pertinence des variables sélectionnées. Cependant, quelques ambiguïtés d'identification subsistent car il faut garder en tête que nous avons travaillé à l'échelle des protéines entières sans avoir eu recours au séquençage par MS/MS ou à la ultra-haute résolution. Ceci est particulièrement vrai dans le cas des sous-types H2A et H2B qui possèdent de très nombreux variants et isoformes ainsi que des profils de modifications complexes. Il peut ainsi arriver que la forme non modifiée d'un variant ait la même masse moléculaire moyenne que la forme modifiée d'un autre variant. Ce problème est très peu présent dans le cas de l'histone H3 et pas du tout dans le cas de l'histone H4 qui ne possède aucun variant.

Cette étude s'intéressant au degré d'acétylation des histones après traitement par un HDACI, nous avons systématiquement associé un incrément de masse de +42 Da à une acétylation. De la même façon, nous avons associé un incrément de masse de +28 Da à une diméthylation ou à deux monométhylations plutôt qu'à une formylation résultant d'un stress oxydant car l'abondance relative de cette modification n'excède pas 0,07%[275]. De manière plus générale, il ne nous a pas été possible de distinguer les espèces isobares à l'échelle des protéines intactes. Pour cela, il nous aurait fallu atteindre les 300 000 de résolution ce qui va bien au-delà des capacités du spectromètre de masse Synapt G2.

À la dose de 1 mM, dix-sept protéoformes différentes ont été identifiées et validées comme étant discriminantes entre les échantillons témoins et les échantillons exposés. Comme le résume le tableau 24, dix de ces dix-sept variables discriminantes correspondent à des formes différentiellement modifiées de l'histone H4. Six correspondent très probablement à l'histone H2B1 et une seule à l'histone H2A1. En observant les ratios d'abondance, les formes les plus acétylées sont surexprimées lors du traitement au butyrate de sodium à 1 mM au détriment des formes les moins acétylées qui apparaissent comme étant sous-exprimées. Les abondances relatives des différentes formes sont comparables puisque la quantité globale d'histones incorporée dans la chromatine est constante. L'exposition au butyrate de sodium semble donc induire une accumulation des formes plus lourdement acétylées d'histones.

Comme résumé tableau 25, la plupart des variables identifiées à la dose de 1 mM l'ont également été à la dose de 2,5 mM mais avec des ratios d'abondance plus élevés. Ces résultats confirment la mise en évidence d'un effet dose-réponse par l'ACP. L'approche histonomique se révèle être suffisamment sensible pour repérer des formes modifiées d'histones dont l'abondance relative ne représente pas plus de 0,1% de toutes les formes modifiées de la protéine. Il est intéressant de noter que les formes discriminantes de l'histone H3 n'ont été révélées qu'à la dose de 2,5 mM, ce qui laisse entendre que H3 serait moins sensible au traitement par le butyrate de sodium que les autres histones de cœur. La comparaison des formes discriminantes en fonction de la dose de butyrate de sodium à l'aide d'un diagramme de Venn (figure 105) montre que dix formes sont communes aux deux doses. En parallèle, sept ont été identifiées exclusivement à la dose de 1 mM et dix à la dose de 2,5 mM.

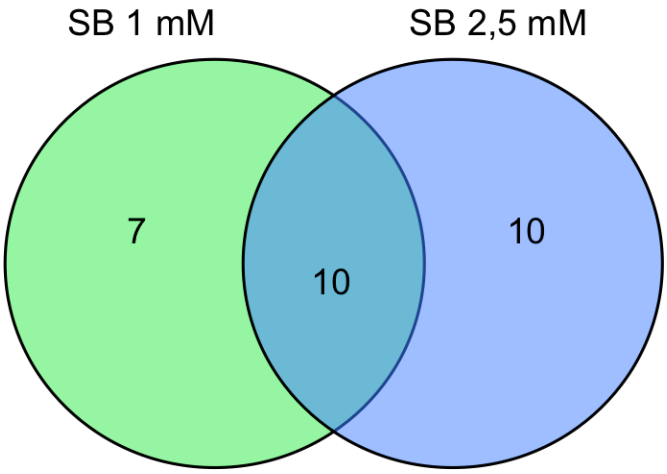


Figure 105 : diagramme de Venn montrant la répartition des formes discriminantes entre les deux classes d'échantillons exposés au butyrate de sodium.

Quelle que soit la dose d'exposition, l'effet le plus marqué se retrouve sur l'histone H4. À elles seules, les différentes formes modifiées de H4 représentent près de 60% des formes discriminantes à 1 mM et 35% à 2,5 mM. Les formes surexprimées portent entre deux et six acétylations, tandis que les formes monoacétylées (11 277,5 Da), et monoacétylées et mono- (11 291,5 Da) ou diméthylées (11 305,5 Da) sont sous-exprimées.

Des travaux récents de Pesavento *et al.*[276] ont montré que dans de très nombreuses lignées cancéreuses humaines, l'acétylation de l'histone H4 se faisait

préférentiellement lorsque la lysine en position 20 était diméthylée (H4K20me2). Ils ont ainsi montré que dans la plupart des cas la forme la plus abondante était la forme monoacétylée sur la sérine N-terminale et diméthylée sur la lysine 20 (H4S1ac-K20me2). Ces affirmations confirment parfaitement ce que nous semblons observer sur les spectres déconvolués de l'histone H4, où la masse moyenne de la forme la plus abondante est 11 305,5 Da, soit un incrément de masse par rapport à la forme non modifiée (11 236,5 Da) correspondant à une acétylation et une diméthylation à l'erreur sur la mesure de masse près. Les variations d'abondance relative des formes modifiées de l'histone H4 entre les conditions d'exposition au butyrate de sodium sont visibles sur les spectres de déconvolution présentés à la figure 106.

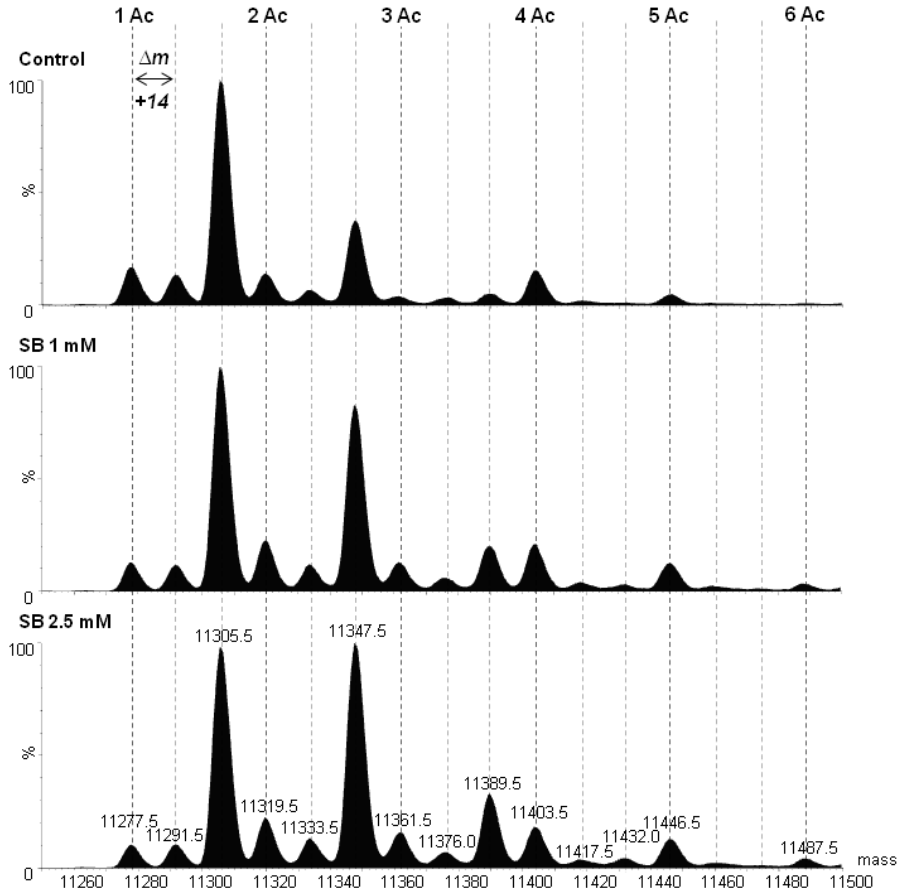


Figure 106 : spectres de déconvolution de l'histone H4 dans un échantillon témoin (*control*) et exposé au butyrate de sodium à 1 ou 2,5 mM.

Les formes identifiées à l'aide des analyses statistiques multivariées sont donc retrouvées manuellement sur les spectres. Les abondances relatives (%) extraites de ces spectres de déconvolution sont présentées figure 107.

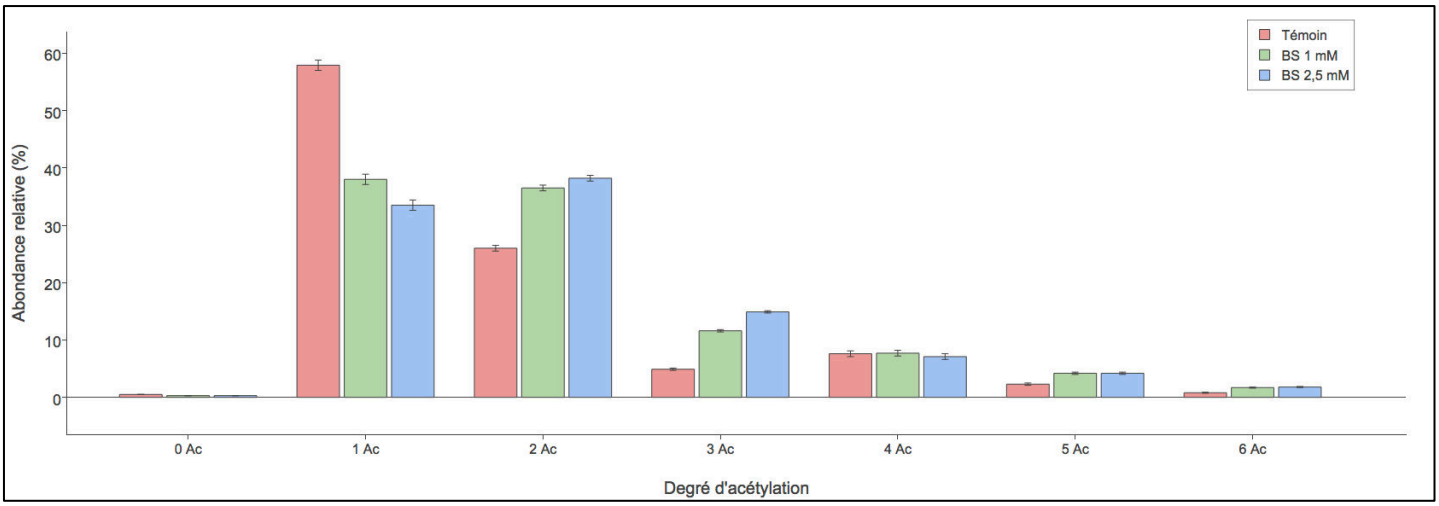


Figure 107 : comparaison des abondances relatives de chaque degré d'acétylation de l'histone H4 dans les échantillons témoins et exposés au butyrate de sodium à 1 ou 2,5 mM. Chaque barre représente la moyenne et l'écart-type des aires sous le pic sur le spectre déconvolué calculés pour trois réplicats biologiques.

Le même travail de validation des résultats à l'échelle des spectres a été fait pour les autres formes d'histone. Concernant H3, quatre protéoformes différentes se sont révélées être sous-exprimées lors d'une exposition au BS à 1 mM et trois surexprimées. Les formes sous-exprimées de H3 sont les formes les moins acétylées et sont éluées aux alentours de 11,5 minutes contre 15,3 minutes pour les formes les plus acétylées qui sont surexprimées. Leur temps de rétention plus court est en accord avec leur plus faible hydrophobicité relative par rapport aux formes hyperacétylées. Comme attendu, les formes non ou monoacétylées de H3 sont sous-exprimées lors d'une exposition au BS quand les formes putativement di- et triacétylées sont, quant à elles, surexprimées.

I.8 Conclusion

Cette partie du travail de thèse avait pour but de prouver l'efficacité et la validité de notre méthode pour le suivi du degré d'acétylation des histones après une exposition à un inhibiteur HDAC dont les effets sur la chromatine sont connus. L'objectif était donc de mettre en évidence le maximum de protéoformes dont l'abondance était affectée par l'exposition. Nous avons ainsi mis en évidence à l'aide de notre approche histonomique l'ensemble des formes d'histones différentiellement acétylées après une exposition au butyrate de sodium. Les modèles statistiques multivariés ont révélé que tous les sous-types d'histone de cœur étaient sensibles à une exposition au butyrate de sodium. Cependant, certaines formes d'histones ne sont concernées qu'à partir d'une dose plus élevée, suggérant ainsi un effet dose réponse. La plupart des formes discriminantes ont pu être identifiées à l'aide des spectres de déconvolution, même si l'identification de certaines espèces isobares reste spéculative et nécessiterait une validation par un séquençage peptidique en aval. En résumé, ces résultats valident la puissance de notre approche histonomique pour cribler de façon rapide et automatique l'ensemble des protéoformes discriminantes entre deux ou plusieurs groupes d'échantillons. Elle pourrait donc s'avérer utile pour le suivi thérapeutique de patients traités par des inhibiteurs HDAC. Ces résultats ont fait l'objet de la publication reproduite en annexe.

II. Exposition à un agent toxique : le benzo[*a*]pyrène

II.1 Introduction

Le benzo[*a*]pyrène (B[*a*]P) est un polluant de la famille des hydrocarbures aromatiques polycycliques (HAP) très répandu dans notre environnement. L'Homme y est quotidiennement exposé *via* l'air qu'il respire ou les aliments qu'il ingère. Les HAP sont générés lors de la combustion incomplète de matières organiques, et les principales sources sont les gaz d'échappement des véhicules diesel, les aliments grillés au charbon de bois, la fumée de cigarette ou encore les rejets industriels dans l'atmosphère[277]. Chez l'Homme, ils sont connus pour leurs effets carcinogènes et reprotoxiques. La base moléculaire de ces effets délétères est imputable principalement au récepteur des arylhydrocarbures (AhR, *aryl hydrocarbon receptor*). Ce récepteur est un facteur de transcription qui appartient à la classe 1 de la famille des protéines bHLH/PAS (*basic Helix-Loop-Helix/PER-ARNT-SIM*). Il est présent dans le cytoplasme des cellules eucaryotes et sa principale fonction après activation par un ligand est la régulation de l'expression des gènes qui codent pour certaines enzymes du métabolisme des xénobiotiques. Il possède plusieurs ligands endogènes mais peut également se lier à certains composés exogènes naturels dont le B[*a*]P.

La liaison du récepteur AhR avec une molécule de B[*a*]P entraîne un changement de sa conformation suivie de sa translocation dans le noyau où il s'associe avec le facteur ARNT (*Aryl Hydrocarbon Nuclear Translocator*) également connu sous le nom de HIF1ß (*Hypoxia Inducible Factor 1ß*). L'hétérodimère AhR/ARNT ainsi formé se lie à des éléments de réponse appelés XRE (*Xenobiotic Responsive Element*) qui sont localisés au niveau des promoteurs des gènes cibles. La plupart de ces gènes cibles codent pour des protéines impliquées dans le métabolisme et l'élimination des xénobiotiques (figure 108).

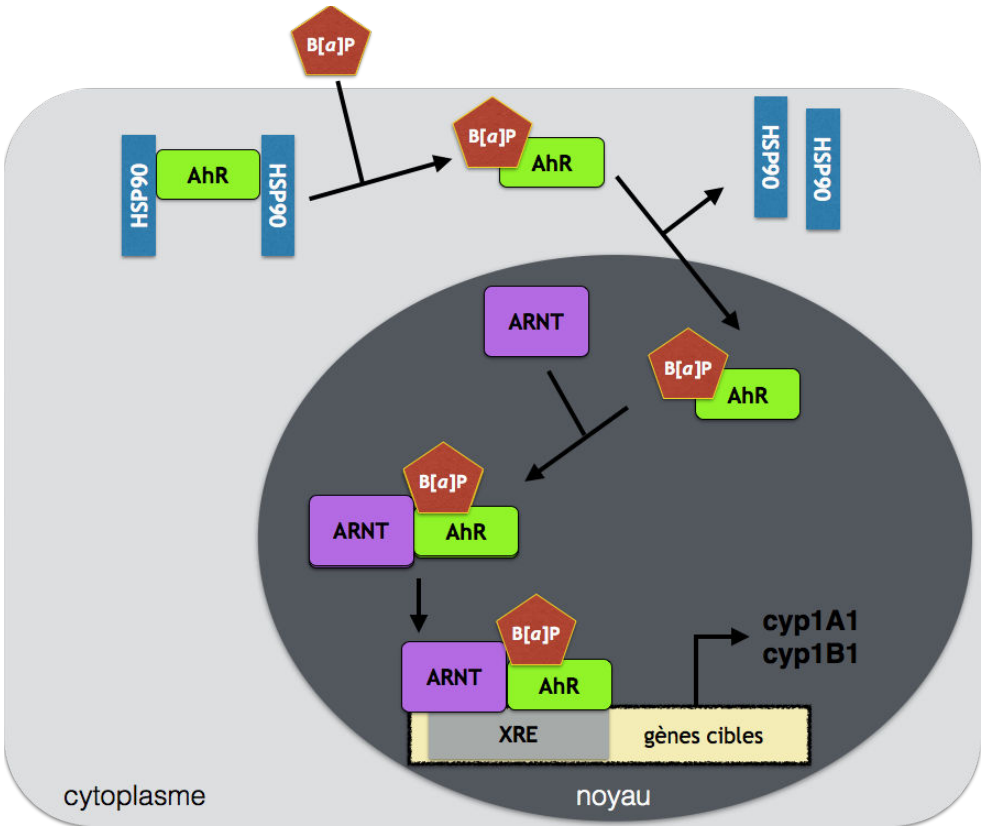


Figure 108 : voie de transduction du AhR après liaison au benzo[*a*]pyrène.

Parmi elles, l'expression des gènes codant pour les cytochromes P450 (CYP) 1A1 et 1B1 et particulièrement augmentée. Ces deux enzymes de phase I sont directement impliquées dans l'oxydation du B[*a*]P en métabolites mutagènes et cancérogènes hautement réactifs, dont le benzo[*a*]pyrène 7,8 diol-9,10 époxyde (BPDE)[278](figure 109). En parallèle, le B[*a*]P peut emprunter d'autres voies minoritaires de métabolisation telles que la voie des sulfo ou glucuronoconjugués qiu aboutissent à la formation de métabolites non toxiques.

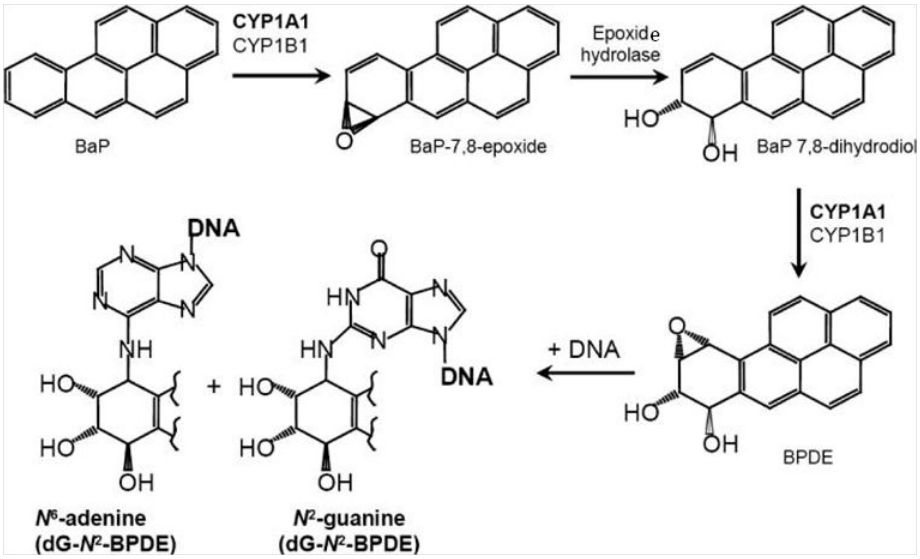


Figure 109 : métabolisme oxydatif du B[*a*]P en BPDE catalysé par les CYP1A1 et 1B1 aboutissant à la formation d'adduits aux bases puriques de l'ADN. D'après[279].

Bien qu'il puisse être directement inactivé par conjugaison avec le glutathion, le BPDE est un métabolite très instable qui induit un stress oxydant et réagit avec n'importe quel type de matériel biologique dont les protéines et l'ADN, avec lesquels il forme des adduits covalents[280]. Les dommages à l'ADN ainsi observés sont de différentes natures. Très récemment, Tung E.W et ses collaborateurs[281] ont montré que le BPDE générait des cassures double brin de l'ADN, dommages qui sont les plus génotoxiques.

De nombreux tissus sont concernés par la survenue de dommages à l'ADN à la suite d'une exposition au B[*a*]P. Parmi eux, le placenta est un véritable organe cible de par le niveau d'expression élevé de AhR[282] et la présence du CYP1A1[283]. Outre un remodelage de l'architecture vasculaire du placenta observé chez des femmes enceintes fumeuses[284], les fonctions endocrines des trophoblastes sont également perturbées par une exposition aux HAP[285]. Certaines études effectuées sur la lignée trophoblastique humaine BeWo ont également prouvé qu'une exposition au B[*a*]P diminuait la sécrétion de la sous-unité ß de l'hormone gonadotrophique chorionique (ß-hCG)[286]. La génotoxicité et la perturbation endocrinienne consécutives à une exposition placentaire au B[*a*]P étant deux phénomènes modulant la structure de la chromatine, nous avons utilisé notre approche histonomique pour tenter de détecter une éventuelle perturbation du code histone associée à ces effets délétères. La révélation de marqueurs

histoniques placentaires d'exposition au B[*a*]P pourrait apporter dans un premier temps une information mécanistique sur son mode d'action toxique à l'échelle placentaire. Ceci permettrait de mieux comprendre et d'anticiper la survenue de certaines pathologies chroniques chez l'adulte en lien avec une exposition *in utero* aux HAP. Cependant, il a été démontré que les modifications de la chromatine à la suite d'une exposition à des polluants environnementaux ne dépendaient pas du polluant lui même mais de son mode d'action. Ainsi, tous les ligands du AhR (B[*a*]P, dioxines, polychlorobiphényls) induiraient les même marques chromatiniennes au niveau des gènes cibles[287]. Nous pouvons donc imaginer obtenir *via* notre approche un profil d'histones caractéristique d'une toxicité AhR dépendante, ce qui permettrait plus largement de classer les placentas après délivrance en fonction de leur historique d'exposition à un certain type de xénobiotiques environnementaux, puis d'en déduire les risques encourus par le nouveau-né.

II.2 Exposition des cellules BeWo au benzo[*a*]pyrène

L'objectif du projet ANR PLACENTOX en relation avec lequel s'est inscrit ce travail était de mieux comprendre la toxicité placentaire du B[*a*]P dans le cadre d'une exposition chronique environnementale de femmes enceintes. La dose choisie pour l'exposition *in vitro* au B[*a*]P a été la même pour l'ensemble des études effectuées par les différents partenaires, que ce soit sur les cellules BeWo ou sur les primo-cultures de trophoblastes. La dose de 1 µM qui a été choisie est très proche des concentrations de B[*a*]P auxquelles l'Homme est exposé quotidiennement à travers l'alimentation[288] ou la fumée de cigarette[289]. Ce choix a également été conforté par les travaux de Le Vee *et al.*[290] qui prouvent que le B[*a*]P perturbe la sécrétion de ß-hCG chez les cellules BeWo lors d'une exposition *in vitro* à des doses de l'ordre de 0,5 à 1 µM. Les cellules BeWo (clone b30) ont donc été exposées en flasque au B[*a*]P à 1 µM pendant 24 heures. En parallèle, un nombre équivalent de flasques témoins a été exposé dans des conditions identiques au véhicule ayant servi à solubiliser le B[*a*]P, à savoir le DMSO.

II.3 Extraction et profilage LC-MS des histones

II.3.1 Dosage des protéines et contrôles

Les 30 extraits histoniques obtenus ont été dosés par la méthode BCA et les résultats pour ceux obtenus après précipitation par le TCA sont présentés au tableau 26. Le nombre de cellules contenu dans chaque flasque est parfois hétérogène, ce qui peut expliquer la différence de concentration observée entre les différents réplicats biologiques.

Tableau 26 : concentrations des différents mélanges d'histones extraits à partir de cellules BeWo témoins (DMSO, vert) et exposées au B[*a*]P à 1 µM (rouge).

	Conc. Histones (µg/µL)
DMSO_1	0,83
DMSO_2	1,31
DMSO_3	1,28
DMSO_4	0,90
DMSO_5	0,86
DMSO_6	1,02
DMSO_7	1,13
DMSO_8	1,18
DMSO_9	0,95
DMSO_10	1,43
DMSO_11	1,45
DMSO_12	1,38
DMSO_13	1,58
DMSO_14	1,56
DMSO_15	1,65
B[*a*]P_1	0,61
B[*a*]P_2	0,78
B[*a*]P_3	0,83
B[*a*]P_4	0,72
B[*a*]P_5	0,64
B[*a*]P_6	0,57
B[*a*]P_7	0,90
B[*a*]P_8	0,72
B[*a*]P_9	0,88

B[*a*]P_10	0,59
B[*a*]P_11	0,42
B[*a*]P_12	0,66
B[*a*]P_13	0,82
B[*a*]P_14	0,63
B[*a*]P_15	0,42

Les profils électrophorétiques de 3 réplicats biologiques par condition d'exposition sont présentés à la figure 110. Pour des quantités de protéines déposées équivalentes (environ 5 µg), tous les profils semblent homogènes et les échantillons ne semblent pas contenir de contaminant majeur.

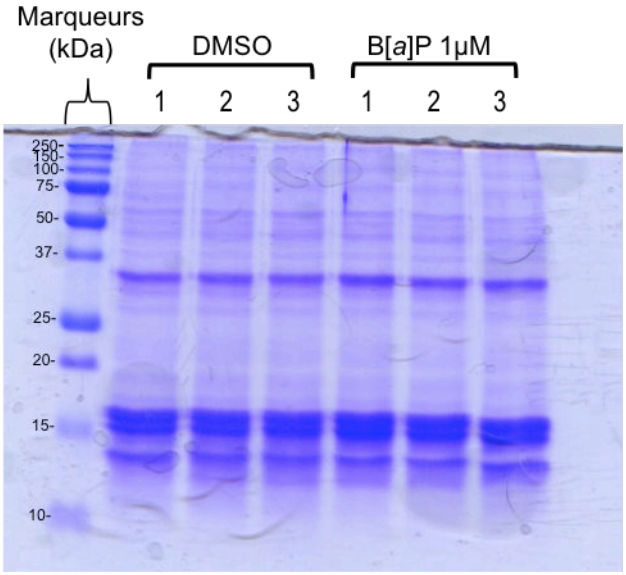


Figure 110 : SDS-PAGE 13% montrant les profils électrophorétiques d'extraits histoniques obtenus à partir de cellules BeWo exposées au DMSO ou au B[*a*]P à 1 µM.

Une concentration cible de 0,3 µg/µL a ensuite été obtenue en diluant chaque échantillon dans de l'eau milliQ contenant 0,05% d'acide formique (solvant A).

II.3.2 Profilage LC-MS des histones extraites

Comme décrit précédemment, l'échantillon QC a été constitué en mélangeant un volume de 3 µL de chaque échantillon préalablement dilué. Une fois l'ordre d'injection orthogonalisé, environ 1,5 µg de chaque échantillon ont été injectés sur la colonne chromatographique, et 5 µL de l'échantillon QC ont été injectés tous les 5 échantillons. La figure 111 présente les chromatogrammes d'un échantillon témoin et d'un échantillon exposé au B[*a*]P à 1 µM. La nature et l'ordre d'élution des différents sous-types d'histones de coeur sont identiques pour chacune des conditions d'exposition.

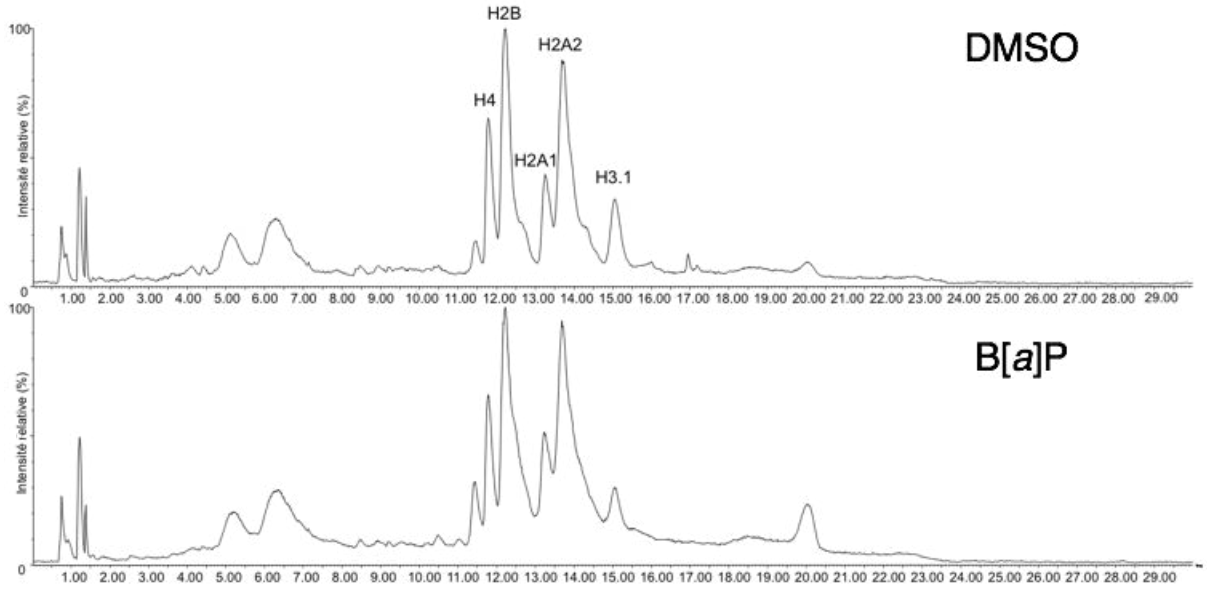


Figure 111 : chromatogrammes d'un échantillon témoin (DMSO) et d'un échantillon exposé au B[*a*]P à 1 µM.

II.4 Prétraitement et normalisation des données

II.4.1 Prétraitement par XCMS

L'ensemble des fichiers convertis au format mzData a été prétraité sous R à l'aide de XCMS. L'étape de réalignement des chromatogrammes a permis de corriger la très légère déviation des temps de rétention observée à la figure 112.

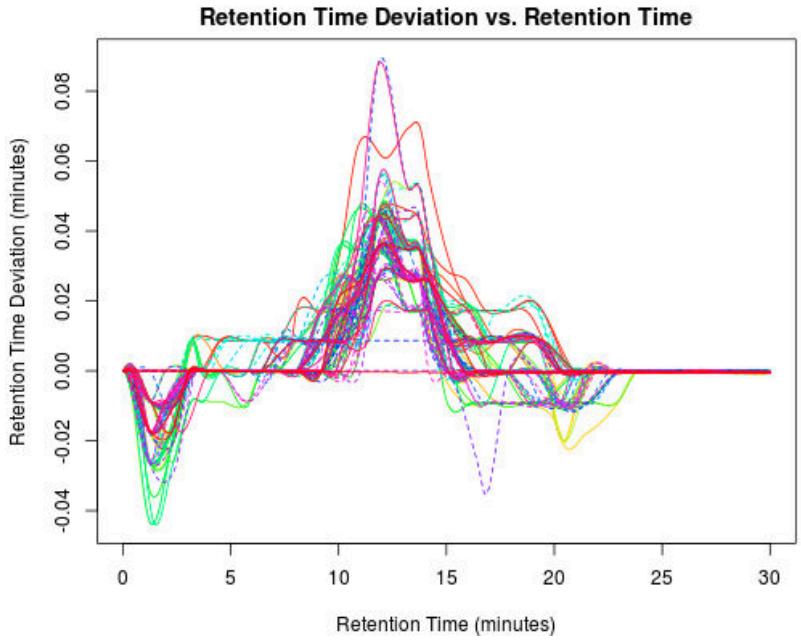


Figure 112 : déviation du temps de rétention observée en fonction du temps de rétention pour l'ensemble des échantillons analysés. Chaque ligne de couleur représente un échantillon.

Les chromatogrammes réalignés par l'algorithme obiwarp sont présentés à la figure 113, où l'on peut observer qu'ils se superposent parfaitement, avec simplement une variation de l'intensité entre certains échantillons.

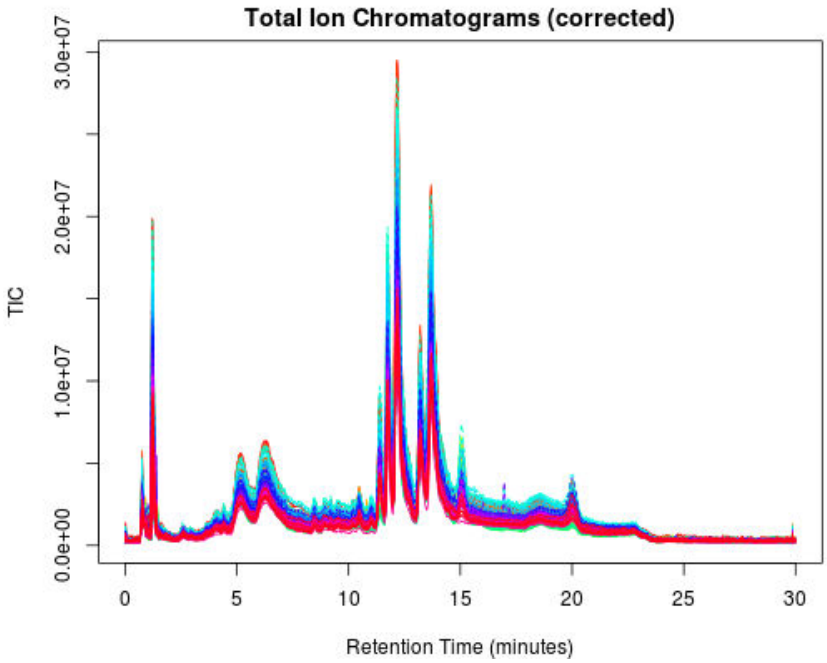


Figure 113 : superposition de l'ensemble des chromatogrammes (TIC) après réalignement et correction des temps de rétention.

Après toutes les étapes de prétraitement, la matrice *X* obtenue contenait 16 095 variables différentes avec chacune un couple t_R_*m/z* unique. Le premier filtre appliqué sur le temps de rétention a permis de réduire le nombre de variables à 12 014. Sur cette série d'échantillons, la médiane des coefficients de variation (CV) était de 18,3%, ce qui représente une précision globale acceptable pour cette analyse.

II.4.2 Normalisation des données

Après normalisation par la médiane, transformation logarithmique et redimensionnement de Pareto, les caractéristiques des différentes variables ont été évaluées à l'aide de boîtes à moustaches. La figure 114 représente ainsi l'effet de ces étapes de normalisation sur les caractéristiques de position de l'intensité de plusieurs variables choisies aléatoirement.

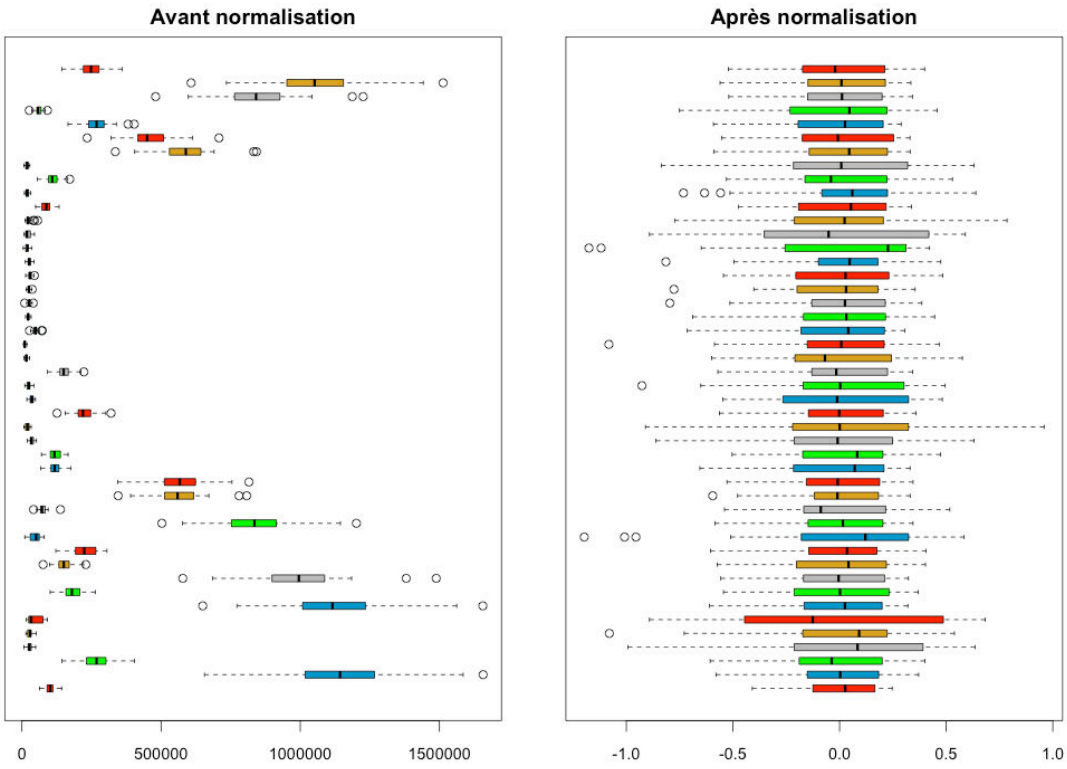


Figure 114 : boîtes à moustaches ou *box plots* résumant les caractéristiques, avant et après normalisation, de 50 variables sélectionnées aléatoirement parmi les 12 014 contenues dans la matrice *X*. L'axe horizontal représente l'intensité des variables.

Nous pouvons observer que ces caractéristiques sont très homogènes après normalisation des données. D'autre part, la densité de probabilité de l'intensité de toutes les variables a été estimée. La figure 115 montre que la normalisation permet d'obtenir une distribution gaussienne de l'intensité des variables.

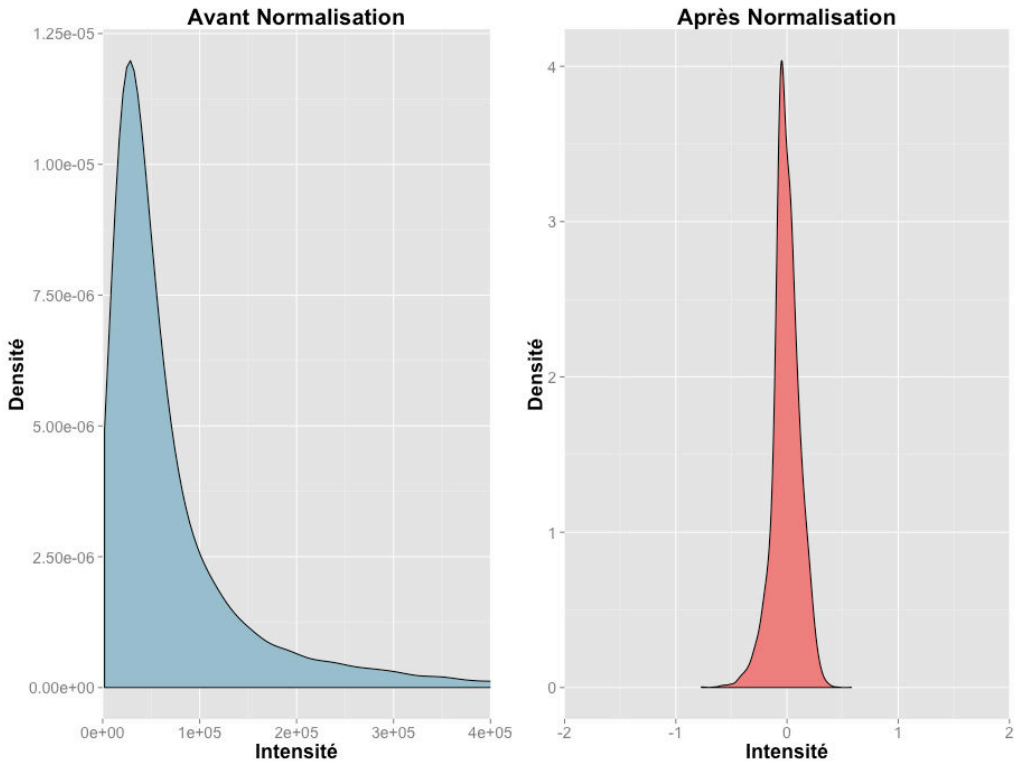


Figure 115 : estimation par noyau de la densité de probabilité de l'intensité des 12 014 variables contenues dans la matrice *X* avant (gauche) et après (droite) normalisation.

II.5 Analyses statistiques descriptives

Avant d'explorer les différentes sources de variabilité au sein des données, cinq échantillons par condition ont été sélectionnés aléatoirement et retirés du jeu de données d'apprentissage pour constituer le jeu de données de prédiction. Contrairement au cas précédent du butyrate de sodium où nous cherchions à lister de manière exhaustive l'ensemble des formes d'histones dont le degré d'acétylation variait avec l'inhibition des HDAC, nous sommes avec le B[*a*]P davantage dans une démarche de recherche de marqueurs spécifiques d'exposition. Nous avons donc cherché à établir une liste plus réduite de marqueurs les plus discriminants possible entre les témoins et les échantillons exposés. Cette logique différente nous amènera à ajuster certaines étapes de l'analyse statistique.

II.5.1 Classification ascendante hiérarchique

La classification ascendante hiérarchique appliquée au jeu de données d'apprentissage permet d'agglomérer naturellement les échantillons selon leur condition d'exposition sur la base de l'intensité des 50 variables les plus significatives. Le contraste entre le profil d'expression de ces 50 variables au sein des deux groupes d'échantillons est visible sur la figure 116. Les deux profils discriminants sont définis par les deux branches les plus longues du dendrogramme.

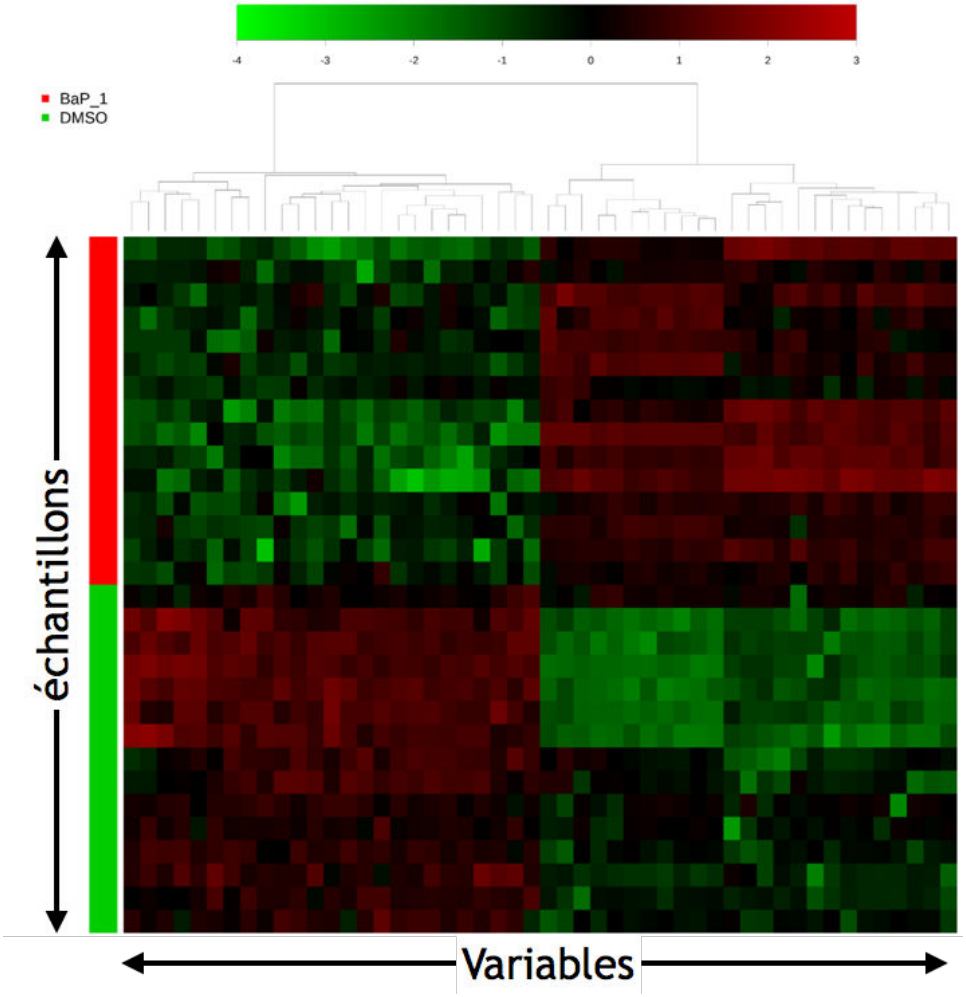


Figure 116 : classification ascendante hiérarchique et représentation *heat map* des 50 variables les plus significativement différentes entre les deux groupes d'échantillons d'après les résultats d'un test ANOVA. L'axe vertical représente les différents réplicats biologiques pour chacun des groupes (rouge = B[*a*]P, vert= DMSO). Le gradient de couleur représente les intensités normalisées des variables sur une échelle logarithmique. Les points de couleur verte traduisent une faible abondance tandis que ceux de couleur rouge représentent une forte abondance des variables.

Toutes ces variables discriminantes ont un temps de rétention compris entre 11 et 13 minutes, ce qui signifie qu'elles correspondent toutes à des formes d'histone H4 ou H2A/H2B non modifiées ou portant différentes modifications post-traductionnelles. En résumé, les profils d'histones des échantillons témoins et des échantillons exposés au B[*a*]P présentent suffisamment de différences pour permettre la classification des échantillons simplement à l'aide de cette analyse non supervisée.

II.5.2 Analyse en composantes principales

Le modèle ACP comporte trois composantes principales, dont deux seulement sont représentées à la figure 117. La somme de ces trois composantes principales résume 24% de la variance totale (R^2X_{cum} = 0,24). La variance expliquée par ce modèle n'est pas très élevée, ce qui ne reflète pas directement la qualité du modèle[291]. En effet, exclure un échantillon aberrant d'un modèle peut par exemple faire baisser la variance expliquée de moitié alors que le modèle en sera d'autant meilleur.

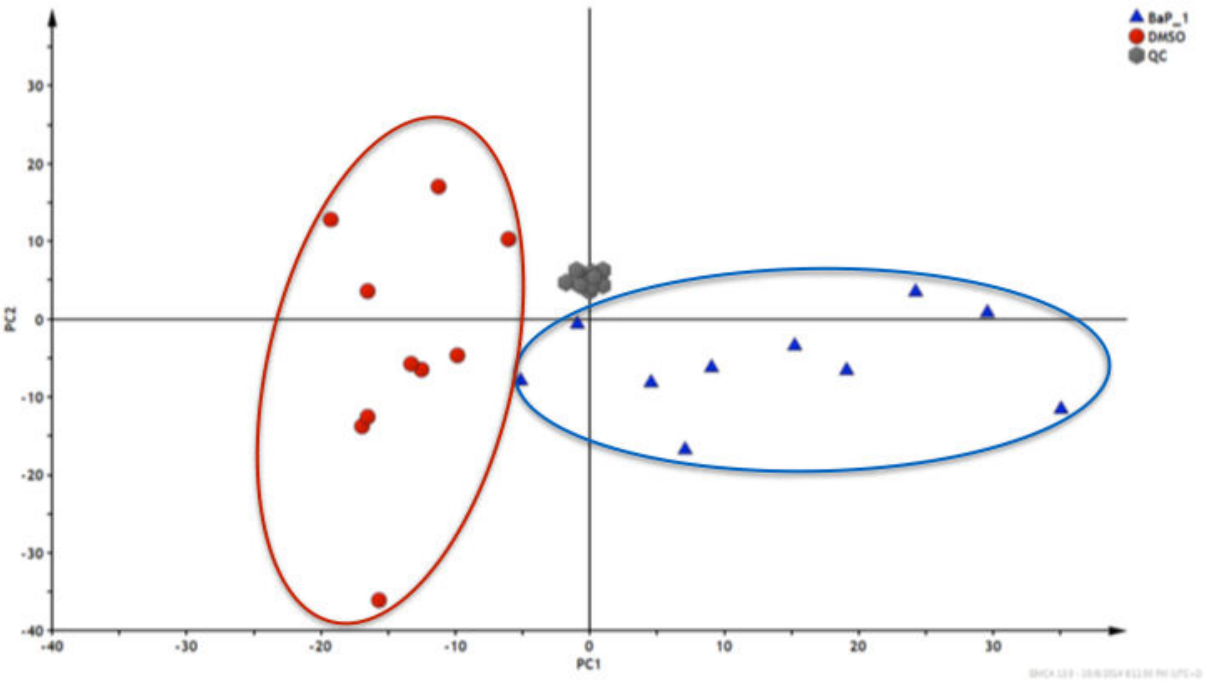


Figure 117 : *scores plot* 2D d'une ACP représentant uniquement les deux premières composantes principales PC1 et PC2. Les trois classes d'échantillons sont représentées par des formes et des couleurs différentes : ○ = témoins, △ = B[a]P 1 µM et ⬢ = QC.

D'après le *scores plot* présenté à la figure 117, l'ACP permet de séparer les échantillons en deux groupes distincts sur la composante principale 1 (PC1). La variabilité maximale entre les échantillons est donc bien liée à l'exposition au B[*a*]P. Nous pouvons également observer que les échantillons sont relativement dispersés au sein d'une même classe, traduisant une certaine hétérogénéité de la réponse à l'exposition. Pourtant, l'exploration de la distance de chaque échantillon par rapport au modèle (DModX) représentée à la figure 118 montre l'absence d'individu aberrant. Enfin, le regroupement des réplicats de l'échantillon QC au centre du *scores plot* atteste de la stabilité de notre analyse.

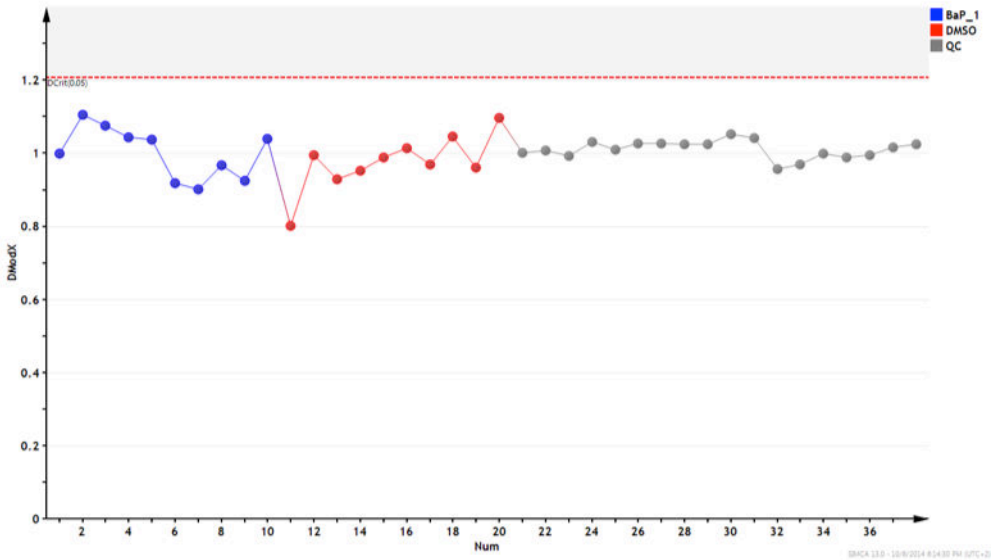


Figure 118 : distance de chacun des individus par rapport à la première composante principale du modèle ACP. Chaque point de couleur représente un échantillon appartenant à l'une des trois classes (rouge = DMSO, bleu = B[*a*]P et gris = QC).

L'exploration du *loadings plot* permet de se faire une première idée de la nature et du nombre de variables responsables de la séparation naturelle des échantillons. La figure 119 sépare donc à gauche les variables plus abondantes dans les échantillons témoins et à droite celles plus abondantes dans les échantillons exposés au B[*a*]P. La séparation entre ces variables se fait selon la PC1.

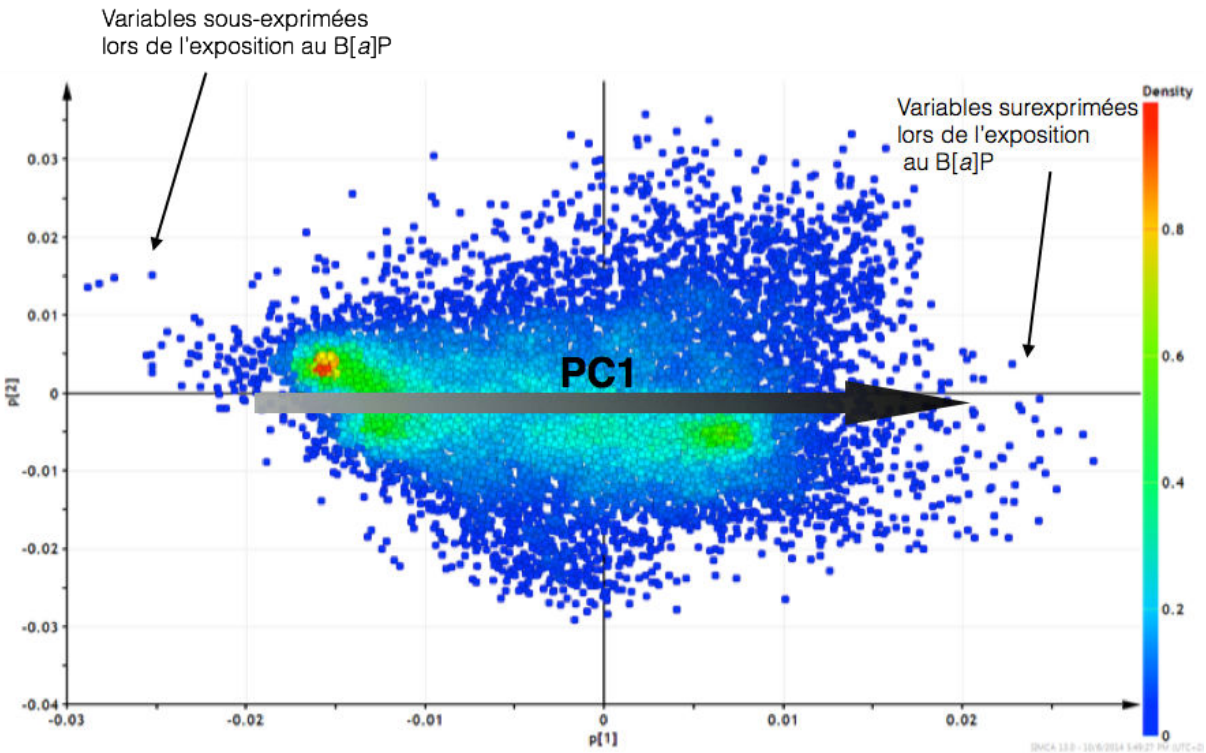


Figure 119 : *loadings plot* du modèle ACP représentant les deux premières composantes. Le gradient de couleur représente la densité des variables lorsqu'elles sont superposées.

Nous pouvons ainsi observer que la majorité des variables forme un nuage de point assez dense au centre du graphique, et que peu de variables semblent réellement se détacher à gauche ou à droite de ce nuage. Dans ces conditions, il est difficile de savoir quelle variable considérer ou non.

II.5.3 Bilan des analyses statistiques non supervisées

La classification ascendante hiérarchique et l'ACP révèlent sans ambiguïté que l'exposition des cellules BeWo pendant 24h au B[*a*]P à 1 µM induit un changement substantiel des profils d'histones. Ce résultat intermédiaire est une première réponse à la question initialement posée : le B[*a*]P est-il à l'origine d'une perturbation du code histone ? Sans même savoir exactement quelles sont les formes d'histones discriminantes, l'approche globale mise au point nous permet d'affirmer que le B[*a*]P perturbe *in vitro* le code histone.

II.6 Classification des échantillons et analyses statistiques prédictives

Les analyses supervisées nous permettent de construire des modèles prédictifs capables de classer les échantillons exposés ou non au B[*a*]P sur la base de leur profil d'histones. Contrairement au cas du butyrate de sodium, nous ne cherchons pas ici à comparer plusieurs doses mais simplement des échantillons exposés ou non. Les deux classes peuvent donc être discriminées directement à l'aide d'une analyse binaire OPLS-DA. Le modèle OPLS-DA a donc été construit à partir du jeu de données d'apprentissage. Comme précédemment, le modèle final présente une composante prédictive et une composante orthogonale (figure 120). Ses caractéristiques statistiques sont les suivantes : R^2Y = 0,99 et Q^2Y = 0,71. Enfin, la p-value fournie par le test ANOVA est de 6,3 x 10^{-4}.

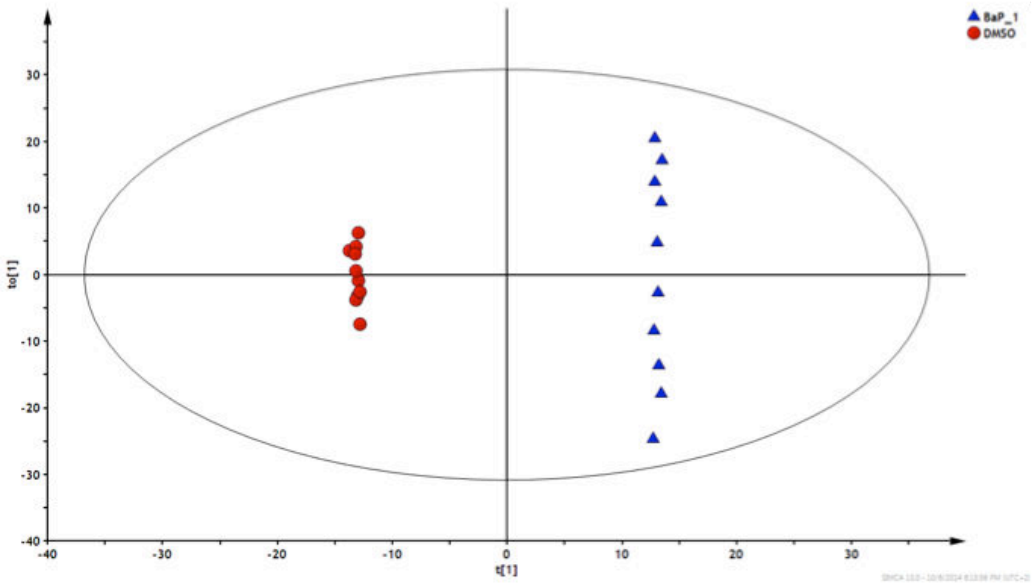


Figure 120 : *scores plot* du modèle OPLS-DA obtenu à partir du jeu de données d'apprentissage contenant les échantillons témoins (DMSO) et les échantillons exposés au B[*a*]P à 1 µM.

Nous pouvons observer sur la figure 120 que la composante prédictive sépare très clairement les échantillons en fonction des conditions d'exposition. Sur la composante orthogonale, les échantillons exposés au DMSO sont relativement groupés, contrairement aux échantillons exposés au B[*a*]P à 1 µM qui sont plus dispersés, laissant supposer qu'il existe une certaine hétérogénéité dans la réponse des cellules BeWo à l'exposition au B[*a*]P.

La projection des échantillons du jeu de données de prédiction dans l'espace défini par ce modèle OPLS-DA confirme que le modèle est valide. Nous pouvons observer sur la figure 121 que la composante prédictive conserve son pouvoir discriminant entre les échantillons témoins et ceux exposés au B[*a*]P.

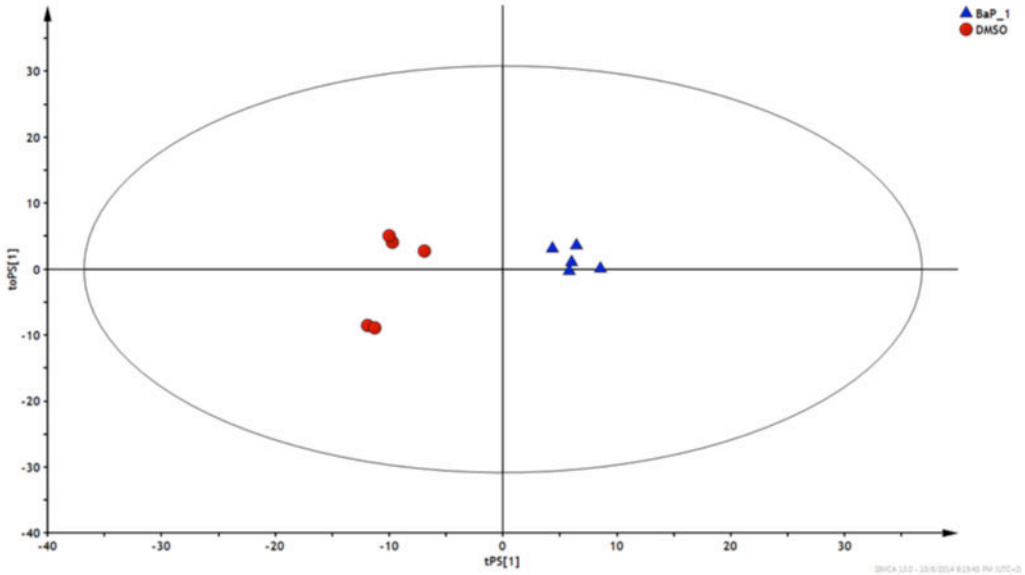


Figure 121 : *scores plot* de la projection dans le modèle OPLS-DA défini précédemment du jeu de données de prédiction contenant les échantillons témoins (DMSO) et les échantillons exposés au B[a]P à 1 µM.

L'examen de la table des erreurs de classification révèle que 100% des échantillons ont été correctement classés par ce modèle (tableau 27).

Tableau 27 : erreurs observées après la classification par le modèle OPLS-DA du jeu de données de prédiction contenant les échantillons témoins (DMSO) et les échantillons exposés au B[a]P à 1 µM.

	Membres	Corrects	DMSO	B[*a*]P 1 µM
DMSO	5	100%	5	0
B[*a*]P 1 µM	5	100%	0	5
Pas de classe	0	-	0	0
Total	10	100%	5	5
Prob. Fisher	0,004			

II.6.1 Formes d'histones discriminantes associées à l'exposition au B[*a*]P

Pour extraire les variables ayant la plus forte contribution dans la discrimination des deux classes d'échantillons, nous avons utilisé la représentation graphique S-Plot™. Le S-Plot™ représente sur l'axe vertical la corrélation de chaque variable et sur l'axe horizontal leur covariance. La corrélation peut être apparentée à la fiabilité de la variable sur la composante prédictive, tandis que la covariance correspond, elle, à sa magnitude au sein des deux classes d'échantillons. La figure 122 présente ces caractéristiques pour chacune des variables contenues dans la matrice *X*.

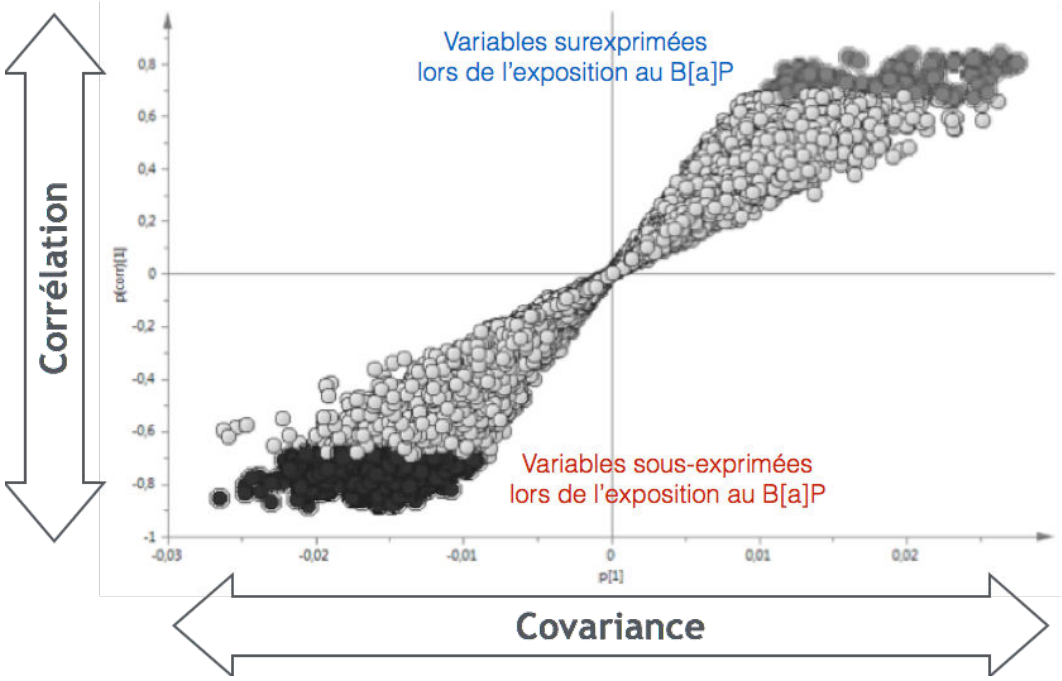


Figure 122 : représentation *S-Plot™* des variables sur la composante prédictive. Les points surlignés en noir et en gris foncé aux extrémités supérieure droite et inférieure gauche correspondent aux variables sélectionnées.

Les variables sont donc sélectionnées en fonction de leur localisation sur ce S-Plot™. Les variables situées dans le coin supérieur droit correspondent aux variables surexprimées lors de l'exposition au B[*a*]P tandis que celles situées dans le coin inférieur gauche correspondent aux variables sous-exprimées lors de l'exposition. Nous avons choisi de ne conserver que les variables dont le coefficient de corrélation absolu était supérieur à 0,6. Nous avons également effectué une

étape de filtration supplémentaire sur le score VIP en prenant 1,5 comme valeur seuil. Ces étapes ont abouti à une liste de 212 variables. Cette liste a davantage été raffinée en appliquant un filtre sur les CV (< 30%) ainsi que sur les *q-value* (< 0,001). Ensuite, l'information redondante liée aux différents états de charge d'une même variable a été supprimée comme précédemment pour aboutir à une liste finale de 18 protéoformes discriminantes présentée dans le tableau 28.

Parmi les variables discriminantes retenues, nous pouvons observer la présence de formes d'histone H2A, H2B et H4. Quelques variables correspondant potentiellement à deux formes modifiées de H3.1 ont également émergé des analyses statistiques. Cependant, seuls un ou deux états de charge par forme étaient significativement différents entre les deux groupes. De plus, lors de l'inspection sur les spectres des pics correspondant, le rapport signal sur bruit était mauvais. Nous avons donc estimé qu'une attribution de ces variables était trop hasardeuse, et par conséquent elles n'ont pas été intégrées à la liste finale des protéoformes discriminantes. Concernant les autres variants probablement identifiés, l'ambiguïté est parfois présente entre deux protéoformes différentes, comme dans le cas de l'exposition au butyrate de sodium. Certaines formes telles que H2A.Z et H4 dont les profils sont moins complexes posent peu de problème d'identification, contrairement à des formes telles que H2A1 et H2B1 dont la masse d'une isoforme non modifiée peut correspondre exactement à la masse d'une autre isoforme modifiée. Ainsi, TagIdent fournit fréquemment deux, voire trois, possibilités d'identification et il nous est impossible de trancher sur la seule base des masses moléculaires moyennes de protéines intactes. Le but de ce travail étant de révéler des marqueurs robustes et spécifiques d'exposition au B[*a*]P, nous avons choisi de nous concentrer sur les variables les plus discriminantes identifiables.

Tableau 28 : identifications les plus probables des variables discriminantes entre les échantillons témoins (DMSO) et les échantillons exposés au B[*a*]P à 1 µM.

Identité	N° d'accès UniProtKB	Masse moyenne théorique de la forme non modifiée (Da)	Masse moyenne observée de la forme non modifiée (Da)	Incrément de masse observé (Da)	PTMs	Score VIP	p(corr)	CV (%)	FDR (q-value)	Rat
H2A.Z	P0C0S5	13 421,5	13 420,5	0	*	2,55	-0,87	8,7	4,90E-08	-0,
				+42	1 ac	2,74	0,72	13,1	1,74E-07	1,
H2A-1B/E	P04908	14 004,3	14 004	0	*	2,30	-0,81	27,1	4,96E-06	-0,
				+14	1 me1	1,61	-0,79	8,7	6,40E-05	-0,
				+28	1 me2 / 2 me1	1,60	-0,88	6,6	1,72E-08	-0,
				+42	1 ac	1,60	-0,85	5,1	3,46E-07	-0,
				+98	2 ac + 1 me1/1 ac + 1 me3 + 1 me1/1 ac + 2 me2/1 ac + 4 me1	1,84	0,64	21	9,45E-05	0,
H2A-2C / H2B-1M	Q16777 / Q99879	13 857,2 / 13 858,0	13 857,5	+14	1 me1	1,76	0,63	17,3	2,06E-05	0,
				+98,5	1 ac + 1 me3/ 1 ac + 1 me2 + 1 me1/ 1 ac + 3 me1	2,33	0,82	9,2	5,52E-07	0,
				+113,5	2 ac + 1 me1/1 ac + 1 me3 + 1 me1/1 ac + 2 me2/1 ac + 4 me1	2,22	0,69	21,3	1,18E-05	0,
H2B-1K / H2B-1C / H2B-1O	O60814 / P62807 / P23527	13 758,9 / 13 774,9 / 13 774,9	13 758,5 / 13 774,0	0	H2B-1K	2,10	-0,81	6,2	1,47E-07	-0,
				+14,5	1 me1 / H2B-1C / H2B -1O	2,14	-0,84	5,4	1,72E-08	-0,
				+28	1 me2 / 2 me1 / H2B-1C/0 + 1 me1	1,82	-0,84	6,9	3,30E-08	-0,
				+42 / +28	1 ac/1 me3/1 me2 + 1 me1/H2B-1C/0 + 1 me2/H2B-1C/0 + 2 me1	1,88	-0,84	10,2	2,71E-08	-0,
				+57	1 ac + 1 me1/ 1 me3 + 1 me 1/2 me2/ H2B-1C/O + 1 ac/H2B-1C/0 + 1 me3/H2B-1C/O + 1 me2 + 1 me1/H2B-1C/O + 3 me1	1,91	-0,85	5,5	1,72E-08	-0,
				+70	1 ac + 2 me1/ 1 ac + 1 me2/1 me3 + 2 me1/1 me3 + 1 me2/2 me2 + 1 me1/ H2B-1C/O + 1 ac + 1 me1/H2B-1C/0 + 1 me3 + 1 me1/H2B-1C/O + 2 me2/H2B-1C/O + 4 me1	2,21	-0,81	13,9	3,23E-07	-0,
H4	P62805	11 236,1	11 236,5	+168	4 ac	2,88	0,80	9,4	1,55E-05	1,
				+211	5 ac	1,70	0,68	21,7	0,00013	0,

Si nous recherchons par exemple les deux protéoformes les plus discriminantes sur la base de leur score VIP et de leur ratio d'abondance, nous remarquons qu'il s'agit de l'histone H4 tetra-acétylée ainsi que du variant H2A.Z monoacétylé. L'exposition au B[*a*]P à 1 µM semble donc induire principalement une hyperacétylation de H4 et de H2A.Z, potentiellement en rapport avec son pouvoir d'inducteur transcriptionnel. Ces résultats pourraient donc être en accord avec les travaux récemment publiés par Draker *et al.*[292]. Ces auteurs ont montré que l'acétylation de H4 ainsi que la présence du variant H2A.Z sous forme non modifiée et monoacétylée étaient à l'origine du recrutement du facteur de transcription Brd2 au niveau des nucléosomes lors d'une activation transcriptionnelle. Ce facteur Brd2 est lui-même associé à une activité HAT dirigée vers H4 et H2A[293]. Ainsi le recrutement de complexes à activité HAT par Brd2 au niveau de certains nucléosomes serait à l'origine de la propagation d'une hyperacétylation des histones H2A et H4. Cette hyperacétylation pourrait donc au final servir de rétrocontrôle positif pour favoriser le recrutement de Brd2 et d'autres facteurs de transcription.

L'acétylation de H4 étant, elle, relativement courante et peu spécifique, nous avons préféré explorer davantage la monoacétylation de H2A.Z observée pour tenter de la relier à un phénomène biologique. Dans leurs travaux publiés en 2012, Valdés-Mora *et al.*[294] affirment que la monoacétylation de H2A.Z est une modification clé associée à une dérégulation épigénétique de l'expression des gènes lors des processus de tumorigenèse, ce qui semble concorder avec les propriétés carcinogènes du B[*a*]P. Un autre phénomène cellulaire permet de relier l'exposition au B[*a*]P avec la monoacétylation de H2A.Z : les cassures double brin de l'ADN causées par le métabolite réactif du B[*a*]P, le BPDE[281]. Dans une revue publiée en 2014, Talbert *et al.*[295] rapportent que les cassures double brin de l'ADN sont des sites d'incorporation majeurs du variant H2A.Z par la sous-unité p400 du complexe de remodelage de la chromatine Tip60. Ils décrivent également que ce complexe Tip60 est responsable de l'acétylation de H2A.Z et de H4 induisant une conformation ouverte de la chromatine nécessaire à la mise en place des systèmes de réparation de l'ADN.

En calculant les aires sous le pic des spectres déconvolués correspondant à la forme non modifiée et à la forme monoacétylée de H2A.Z, nous avons pu confirmer

que l'abondance relative de la forme acétylée augmente après l'exposition au B[*a*]P au détriment de la forme non modifiée.

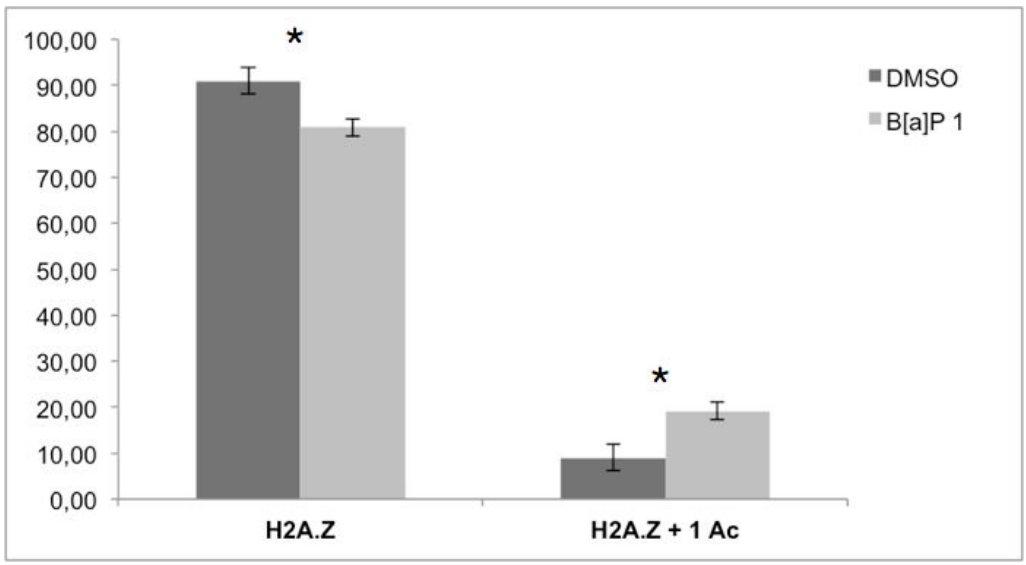


Figure 123 : comparaison des abondances relatives des formes non acétylée et acétylée du variant H2A.Z dans les échantillons témoins et exposés au B[*a*]P à 1 µM. Chaque barre représente la moyenne et l'écart-type des aires sous le pic sur le spectre déconvolué calculés pour trois réplicats biologiques (* : *p-value* < 0,05 avec test de Mann-Whitney).

La monoacétylation du variant H2A.Z semble donc être un marqueur d'exposition au B[*a*]P intéressant d'un point de vue biologique puisqu'il reflète le remodelage de la chromatine qui découle de ses effets toxiques. Nous avons donc cherché à vérifier s'il était possible de discriminer les deux classes d'échantillons sur la seule base de l'abondance relative de la forme monoacétylée de H2A.Z. Pour cela, nous avons utilisé une courbe ROC (*Receiver Operating Characteristic*) qui nous permet de mesurer les performances d'un paramètre en tant que classificateur binaire[296]. Dans ce cas, tous les états de charge détectés pour la forme monoacétylée de H2A.Z ont été regroupés pour former le paramètre évalué. L'examen de l'aire sous la courbe (*Area Under Curve*, AUC) ROC présentée figure 124 permet d'évaluer la performance de la forme H2A.Z monoacétylée en tant que test diagnostique de l'exposition au B[*a*]P.

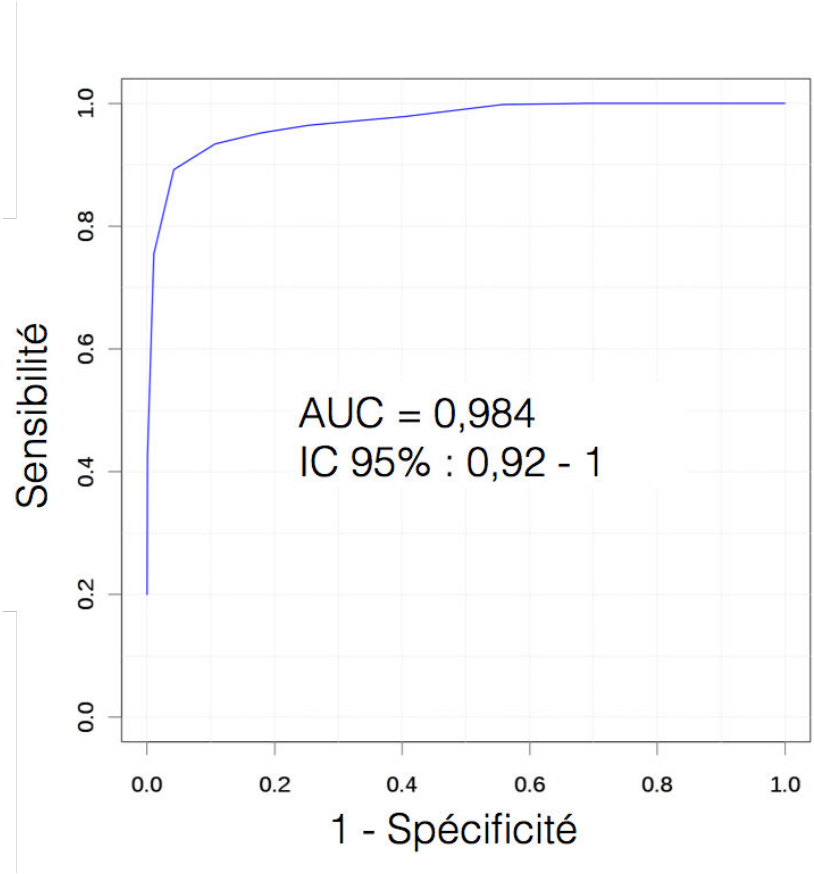


Figure 124 : courbe ROC d'estimation de la performance de la forme H2A.Z monoacétylée comme classificateur binaire. Le taux de vrais positifs (sensibilité) est représenté en fonction du taux de faux positifs (1 – spécificité).

L'AUC permet d'évaluer directement l'intérêt diagnostique du test étudié. Un test présentant une AUC égale à 0,5 est considéré comme nul, tandis que celui présentant une AUC égale à 1 est considéré comme parfait[297]. L'AUC de la courbe ROC fournie ici est de 0,984 et se situe dans la fourchette des tests jugés très informatifs ($0,9 \leq AUC < 1$). L'AUC de 0,984 signifie que dans 98,4% des cas un échantillon exposé au B[*a*]P à 1 µM aura une abondance relative de la forme H2A.Z monoacétylée significativement supérieure à celle d'un échantillon témoin. Ce test possède donc un réel pouvoir discriminant sur la seule base de l'abondance relative de H2A.Z monoacétylée.

II.7 Conclusion

Les analyses statistiques non supervisées révèlent que l'exposition *in vitro* des cellules BeWo au B[a]P à 1 µM induit un changement significatif de leur profil d'histones par rapport aux échantillons témoins. Les profils d'histones des deux classes d'échantillons sont suffisamment différents pour permettre une discrimination naturelle des groupes. L'analyse statistique supervisée OPLS-DA nous a permis de construire un modèle prédictif affichant une précision de classification et de prédiction de 100% laissant envisager une application pratique de ce modèle. Enfin l'extraction des variables les plus significatives a permis l'identification de 18 protéoformes discriminantes parmi lesquelles les deux formes les plus discriminantes, à savoir la forme tetra-acétylée de H4 et la forme monoacétylée de H2A.Z, ont été considérées comme les meilleurs marqueurs de l'exposition au B[*a*]P. L'acétylation de H4 n'étant pas spécifique, nous nous sommes concentrés sur le variant H2A.Z et sa forme acétylée. A l'aide de la littérature, nous avons relié l'incorporation et l'acétylation du variant H2A.Z à différents phénomènes cellulaires en rapport avec les dommages à l'ADN et l'activation transcriptionnelle classiquement induites par le B[*a*]P. Enfin, nous avons montré que l'abondance relative de H2A.Z monoacétylée permettait seule de discriminer les échantillons exposés ou non avec 98,4% de précision. La forme monoacétylée du variant H2A.Z représente donc un marqueur d'exposition au B[*a*]P intéressant qui permet de détecter la survenue d'effets toxiques à l'échelle épigénétique. En accord avec les travaux d'Ovesen *et al.*[287], nous pouvons nous interroger sur la spécificité de ce marqueur par rapport à l'exposition au B[*a*]P. En effet, il est probable que d'autres toxiques ayant les mêmes propriétés génotoxiques et se liant au AhR seraient susceptibles d'induire le même remodelage de la chromatine et donc de présenter le même profil d'histones. Il est donc plus raisonnable d'affirmer que la forme monoacétylée de H2A.Z est un marqueur spécifique non pas du B[*a*]P mais des xénobiotiques carcinogènes génotoxiques en général. Ceci justifierait une étude comparative des effets sur le code histone du B[*a*]P et d'autres xénobiotiques génotoxiques dans le but d'évaluer le caractère spécifique ou non de ce marqueur biologique.

Partie 4

CONCLUSION GÉNÉRALE

Dans le cadre de ces travaux de thèse, nous nous sommes intéressés à l'impact d'une exposition à des xénobiotiques sur la régulation épigénétique, et plus particulièrement sur le code histone. Longtemps sous-estimée, la régulation épigénétique s'avère en effet jouer un rôle clé dans différents mécanismes toxiques. Parmi les mécanismes perturbés lors d'une exposition à des xénobiotiques, le code histone est, de par sa dynamique, une cible de choix. Faisant le lien entre le génome et l'environnement, il intègre l'ensemble des signaux environnementaux et les répercute en modulant l'expression des gènes de manière temps- et tissu- dépendante.

Le développement de certaines maladies chroniques à l'âge adulte a été directement relié à une exposition *in utero* à certains toxiques environnementaux. Dans le cadre du projet PLACENTOX au sein duquel s'est inscrit cette thèse, l'objectif a été de mieux comprendre les mécanismes toxiques à l'origine des effets délétères observés chez le fœtus et chez l'adulte en devenir. Le placenta étant, au-delà d'une simple barrière materno-fœtale, un véritable organe possédant des capacités métaboliques et des propriétés endocrines spécifiques, il est le garant du maintien et du bon déroulement de la gestation. En considérant le placenta comme une cible à part entière pour les xénobiotiques, nous avons cherché à étudier l'impact d'une exposition environnementale à l'échelle du code histone.

Le code histone est, par nature, extrêmement complexe. Englobant à la fois les sous-types, les variants et les isoformes d'histones modifiées post-traductionnellement, c'est plusieurs dizaines de milliers de combinaisons qui peuvent théoriquement être rencontrées. Sa complexité, son hétérogénéité, son dynamisme mais son aspect combinatoire représentent un véritable défi analytique lorsqu'il s'agit de le caractériser et obligent souvent à l'étudier par morceaux plutôt que dans sa globalité. Les méthodes analytiques modernes faisant appel à la chromatographie liquide couplée à la spectrométrie de masse offrent une résolution, une sensibilité et une spécificité suffisantes pour relever ce défi. Différentes stratégies issues de la protéomique ont ainsi vu le jour et permettent de caractériser le code histone à l'échelle des protéines entières ou des peptides. Ces stratégies restent néanmoins exigeantes en terme de préparation d'échantillon

et nécessitent la plupart du temps d'avoir recours à des instruments très résolutifs possédant les modes de fragmentation adéquats.

Dans un contexte de recherche de marqueurs histoniques d'exposition placentaire à des xénobiotiques environnementaux, nous avons mis au point une approche alternative et complémentaire de celles déjà existantes. Cette approche histonomique globale se place en amont d'une caractérisation fine des sites de modification des histones et permet de cribler l'ensemble des protéoformes constitutives du code histone. En partant de cellules BeWo en culture, nous avons mis au point les différentes étapes expérimentales d'une stratégie globale allant de l'extraction des histones à leur analyse par UPLC-ESI-QTOF. Les étapes de traitement des données LC-MS appliquées nous ont permis de supprimer au maximum les sources de variabilité indésirables présentes sur les spectres et inhérentes à ce type d'expérimentation. Enfin, l'utilisation de méthodes statistiques multivariées pour explorer et modéliser la variabilité existant entre les profils d'histones des échantillons a permis de les discriminer en fonction de leur condition d'exposition. Les variables discriminantes ont été identifiées et validées à l'échelle des spectres au moyen de tests statistiques univariés.

Cette approche histonomique validée à l'aide d'un modulateur épigénétique connu, le butyrate de sodium, a ensuite été appliquée à un cas d'exposition à un xénobiotique environnemental qui peut être quotidiennement rencontré, avec un impact non négligeable chez la femme enceinte : le benzo[*a*]pyrène. Dans le cas du butyrate de sodium, inhibiteur de faible spécificité d'histones désacétylases, l'utilisation de notre approche après exposition de cellules BeWo à deux doses différentes a permis d'identifier plusieurs formes d'histones différentiellement acétylées responsables de la discrimination entre les classes. Les modèles statistiques ont également mis à jour une acétylation dose dépendante des histones.

Pour appliquer notre méthode à un cas d'exposition à un polluant environnemental, nous avons choisi le benzo[*a*]pyrène (B[*a*]P), chef de file des hydrocarbures aromatiques polycycliques (HAP). A travers son métabolite réactif, le BPDE, il est connu pour être un génotoxique carcinogène capable d'induire des cassures double brin de l'ADN, mais également une modulation de l'activité

transcriptionnelle de certains gènes. Après une exposition de cellules BeWo au B[*a*]P à 1 µM, notre approche a principalement révélé une augmentation de l'acétylation de H4 et de la monoacétylation du variant H2A.Z. Ces deux marqueurs ont clairement été reliés aux phénomènes cellulaires toxiques induits par l'exposition au B[*a*]P. L'abondance relative du variant H2A.Z s'est révélée être suffisamment spécifique à elle seule pour discriminer correctement les échantillons exposés et les échantillons témoins dans plus de 98% des cas. Cependant, ce marqueur semble davantage être spécifique de la survenue d'un remodelage de la chromatine à la suite de dommages à l'ADN que de l'exposition au B[*a*]P en particulier.

Notre approche histonomique présente donc deux inconvénients. Le premier concerne l'incertitude quant à l'identification de certaines protéoformes sur la base de leur masse moléculaire moyenne. Pourtant, cet inconvénient n'en est pas réellement un. L'approche que nous avons présentée n'a pas pour vocation de se substituer aux approches classiques plus ciblées mais offre une information complémentaire plus globale permettant de repérer les protéoformes sensibles à une exposition à un xénobiotique. L'utilisation en aval de méthodes basées sur la protéolyse enzymatique et l'analyse par spectrométrie de masse en tandem des peptides générés reste donc incontournable pour confirmer et affiner certaines identifications. Elle représentera d'ailleurs la prochaine étape de ce travail.

D'un point de vue biologique, les résultats obtenus par le biais de cette approche histonomique doivent être interprétés avec précaution. La régulation épigénétique étant un mécanisme extrêmement complexe, il est difficile de distinguer une adaptation physiologique d'une adaptation toxique, sans compter celles qui sont fugaces ou trop peu abondantes pour être détectées. En termes de marqueurs révélés, ils semblent davantage être spécifiques d'un type d'événement cellulaire (cassure double brin, activation transcriptionnelle) que d'un polluant en particulier. Ainsi, un travail futur pourra consister à comparer les profils d'histones après exposition à différents polluants appartenant à différentes familles de toxiques ayant un mécanisme de toxicité similaire.

La faible quantité de matériel consommé, la réduction des étapes de préparation d'échantillon, la complémentarité de l'information qu'elle offre ainsi

que sa relative simplicité par rapport aux méthodes conventionnelles font de cette approche globale très directe un réel outil qui devrait trouver toute sa place parmi l'arsenal analytique disponible pour déchiffrer le code histone dans un contexte pathologique ou toxicologique.

PARTIE EXPÉRIMENTALE

1. Culture cellulaire :

1.1. Protocoles

1.1.1. Décongélation d'un cryotube contenant environ 1 million de cellules dans du milieu de culture contenant 10% de DMSO :

⇒ Décongélation rapide de l'ampoule au bain-marie (< 3 min)
⇒ Transfert de la suspension cellulaire dans 10 mL de milieu de culture F-12K contenant 10% de sérum de veau fœtal (SVF)
⇒ Centrifugation 5 min à 200g
⇒ Elimination du surnageant puis reprise du culot dans 5 mL de milieu à 10% SVF
⇒ Transfert dans une flasque T75 (Corning, réf. 10-126-31)
⇒ Ajout de 10 mL de milieu de culture à 10% SVF

1.1.2. Passage des cellules à confluence :

⇒ Rinçage du tapis cellulaire par 2 mL de trypsine-EDTA (Invitrogen, réf. 25300054) puis élimination
⇒ Incubation des flasques avec 2 mL de trypsine-EDTA pendant 5 min à 37°C (jusqu'à détachement des cellules)
⇒ Arrêt de la réaction par dilution de la trypsine avec 3 mL de milieu contenant du sérum de veau fœtal, de la pénicilline et la streptomycine (milieu complet)
⇒ Centrifugation 5 min à 200g
⇒ Elimination du surnageant et reprise du culot dans 10 mL de milieu complet
⇒ Transfert de 2 mL de suspension cellulaire dans chaque nouvelle flasque
⇒ Ajout de 13 mL de milieu complet dans chacune des flasques
⇒ Changement du milieu tous les 2 jours

2. Extraction des histones

2.1. Réactifs

La poudre d'acide 4-(2-hydroxyéthyl)-1-pipérazine éthane sulfonique (HEPES, réf. H3375), le dithiothréitol (DTT, réf. D0632), le chlorure de magnésium ($MgCl_2$, réf.

M2670), le chlorure de potassium (KCl, réf. P9333), le sucrose (réf. S9378), le butyrate de sodium (réf. B5887), l'acide éthylènediaminetétraacétique (EDTA, réf. E6758), l'acide sulfurique (H_2SO_4, réf. 339741) et l'acide trichloroacétique 6,1 N (TCA, réf. T0699) ont été achetés chez Sigma-Aldrich. Le NP-40 (cat. n°11332473001), les inhibiteurs de protéases (cOmplete ULTRA Mini Protease Inhibitor Cocktail, cat. n°04693159001) et les inhibiteurs de phosphatases (cOmplete Phosphatase Inhibitor Cocktail, cat. n°04693116001) proviennent de la société Roche Life Science.

2.2. Tampons

Les différents tampons et solutions utilisés pour l'extraction des histones étaient préparés extemporanément et conservés à 4°C.

Composition des tampons :

Réactifs	**Tampon 1**	**Tampon 2**	**Tampon 3**
HEPES	20 mM	20 mM	20 mM
KCl	10 mM	10 mM	0
$MgCl_2$	1,5 mM	1,5 mM	1,5 mM
DTT	1 mM	1 mM	1 mM
Sucrose	0,24 M	0,24 M	0
EDTA	0	0	2 mM
NP-40	0,15 %	0	0,05 %
Butyrate de sodium	10 mM	10 mM	10 mM
Antiprotéases	1x concentré	1x concentré	1x concentré
Antiphosphatases	1x concentré	1x concentré	1x concentré

2.3. Protocole

⇒ Laisser reposer les culots cellulaires à température ambiante quelques minutes

⇒ Centrifuger les culots secs 3 min à 300g et éliminer le surnageant éventuel
⇒ Rincer les culots par 150 µL de tampon 2
⇒ Centrifuger 3 min à 300g et éliminer le surnageant
⇒ Pour lyser les membranes cellulaires, ajouter 480 µL de tampon 1 et vortexer
⇒ Placer 20 min à 4°C et sous agitation
⇒ Après 5 min d'incubation, homogénéiser à la pipette
⇒ Homogénéiser ensuite fréquemment à la seringue de 1 mL (aiguilles 23G)
⇒ Replacer 5 min à 4°C sous agitation (ne pas dépasser 30 min d'incubation au total)
⇒ Centrifuger 10 min à 3000g
⇒ Prélever les surnageants et les conserver (cytoplasmes)
⇒ Rincer les culots par 180 µL de tampon 2 et centrifuger 10 min à 3000g (X2)
⇒ Pour lyser les membranes nucléaires, ajouter 92 µL de tampon 3 puis 7,5 µL de KCl 4,03 M goutte à goutte (concentration finale KCl = 0,3 M).
⇒ Incuber 30 min à 4°C sous agitation
⇒ Homogénéiser à la pipette, puis à la seringue
⇒ Centrifuger 15 min à 16000g et prélever les surnageants (nucléoplasmes)
⇒ Pour l'extraction acide des histones, resuspendre les culots dans 400 µL d'une solution H_2SO_4 0,4 N et homogénéiser à la seringue
⇒ Incuber 4h à 4°C sous agitation vive en homogénéisant toutes les heures
⇒ Centrifuger 15 min à 16000g et conserver le surnageant (histones)
⇒ Diviser le surnageant d'extraction en deux volumes équivalents
⇒ Pour la première moitié du surnageant, dialyser et concentrer l'échantillon à l'aide d'une unité de filtration AMICON Ultra-0,5 mL (Merck Millipore, cat. n°UFC500396)
⇒ Diluer le filtrat dans 150 µL d'HEPES 20 mM
⇒ Pour la seconde moitié du surnageant, précipiter les histones par du TCA à une concentration finale de 25% à 4°C sur la nuit
⇒ Centrifuger 20 min à 16000g
⇒ Rincer les culots par l'acétone pendant 30 min puis centrifuger 20 min à 16000g (X2)
⇒ Reprendre les culots dans 150 µL d'HEPES 20 mM

3. Contrôles des extraits histoniques

3.1. Dosage des protéines

3.1.1. Dosage des extraits cytoplasmiques et nucléoplasmiques par la méthode de Bradford

Pour le dosage par la méthode de Bradford, nous avons réalisé une gamme d'étalonnage à l'aide de la BSA (Bio-Rad, cat. n°500-0206-MSDS) qui s'étendait de 0 à 0,2 mg/mL. Le réactif de Coomassie utilisé a été acheté auprès de la société Bio-Rad (cat. n°161-0436). La lecture des absorbances s'est faite sur un spectrophotomètre EasySpec (Safas) à la longueur d'onde 595 nm.

3.1.2. Dosage des extraits histoniques par la méthode BCA

Le dosage par la méthode BCA a été réalisé à l'aide d'un kit Pierce® BCA Protein Assay Kit (Thermo Scientific, réf. 23227). La lecture des absorbances s'est faite à une longueur d'onde de 562 nm.

3.2. SDS-PAGE

3.2.1. Réactifs

Tous les pourcentages sont exprimés en rapports masse/volume. Le tampon 2-amino-2-hydroxyméthyl-1,3-propanediol (Tris, cat. n°161-0716), le dodécylsulfate de sodium (SDS) à 10% (cat. n°161-0416), la solution d'acrylamide:bis-acrylamide (cat. n°161-0156) et le persulfate d'ammonium (cat. n°161-0700) ont été obtenus auprès de la société Bio-Rad. Le tétraméthyléthylènediamine (Temed, réf. T9281) a été acheté chez Sigma-Aldrich.

3.2.2. Préparation des gels

Composition des gels de concentration (*stacking*) et de résolution (*resolving*) constitutifs d'un SDS-PAGE 13% :

Réactifs	Gel de résolution (*resolving*)	Gel de concentration (*stacking*)
Tampon Tris-HCl 1,5 M (pH 8,8)	0,375 M	0
Tampon Tris-HCl 0,5 M (pH 6,8)	0	0,125 M
Dodécylsulfate de sodium (SDS)	0,1 %	0,1 %
Persulfate d'ammonium (APS)	0,1 %	0,075 %
Acrylamide:Bis-acrylamide (29:1)	13 %	4 %
Tétraméthyléthylènediamine (TEMED)	0,1 %	0,1 %

3.2.3. Migration

Le système Mini-PROTEAN Tetra Cell (cat. n°164-5052) ainsi que le générateur PowerPac HC de chez Bio-Rad (cat. n°165-8004) ont été utilisés pour la migration des gels. Le tampon de migration est composé de glycine à 1,98 M, de Tris à 250 mM et de 1 % de SDS (poids/volume). Les dépôts sont focalisés sur le gel de concentration pendant 10 min à 120 V puis la migration est réalisée à 200 V jusqu'à ce que le front de migration arrive à l'extrémité inférieure du gel. Les marqueurs de masses moléculaires utilisés provenaient de chez Bio-Rad (Precision Plus Protein Standards, cat. n°161-0373) et couvraient une gamme de masse allant de 10 à 250 kDa.

3.2.4. Coloration

Après migration, les gels sont rincés à l'eau milliQ puis incubés à température ambiante pendant 1h avec le réactif Imperial™ Protein Stain (Thermo Scientific, cat. n°24615). Les gels sont ensuite décolorés par lavages successifs dans l'eau milliQ, jusqu'à obtenir la décoloration souhaitée.

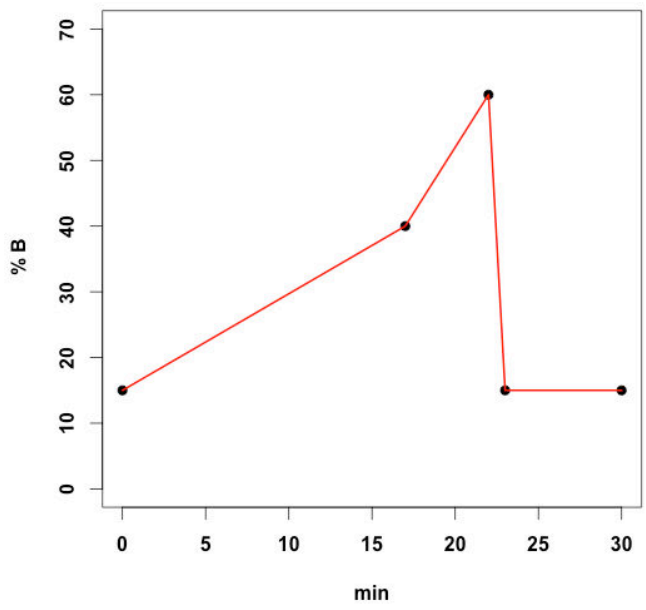
3.3. MALDI-TOF

L'acide alpha-cyano-4-hydroxycinnamique a été utilisé en tant que matrice pour l'analyse des protéines intactes en MALDI-TOF. Pour le dépôt, 1 µL de solution saturée de matrice à 5 mg/mL dans 70% acétonitrile / 0,1% acide trifluoroacétique (TFA) a été mélangé avec 1 µL d'échantillon. La totalité a été déposée sur la plaque MALDI en inox et laissée à l'air libre jusqu'à séchage du dépôt.

Le spectromètre de masse Voyager-DE™ Pro de chez Applied Biosystems a été utilisé en mode linéaire.

4. Séparation chromatographique Acquity UPLC

Conditions chromatographiques :

Colonne	Acquity UPLC BEH C_{18} 2,1*150 mm, 1,7 µm, 300 Å (Waters, cat. n°186003687)
Phases mobiles	A : 0,05 % acide formique dans H_2O B : 0,05% acide formique dans ACN
Débit	0,3 mL/min
Gradient	
Température de colonne	+55°C
Mode d'injection	*Partial Loop*
Température du passeur	+4°C

5. Spectromètre de masse Synapt G2 HDMS

Plan factoriel complet pour le criblage des paramètres de source :

Pour l'interprétation du plan factoriel complet réalisé, les effets principaux *E* de chaque facteur *k* ont été calculés aux bornes supérieures et inférieures à partir des réponses *y(+)* et *y(-)* mesurées sur les *n* essais d'après les formules suivantes :

$$E(k_+) = \sum_{i=1}^{i=n/2} y_i(+)$$

$$E(k_-) = \sum_{i=1}^{i=n/2} y_i(-)$$

Les effets moyens *E* de chaque facteur *k* ont été calculés à partir des réponses *y* mesurées sur les *n* essais d'après la formule suivante :

$$E(k) = \frac{\sum_{i=1}^{i=n/2} y_i(+) - \sum_{i=1}^{i=n/2} y_i(-)}{n/2}$$

Les paramètres de la méthode finale sont présentés ci-dessous :

Paramètre	Valeur
Polarité	ESI positif
Mode	Résolution (simple réfléctron V)
Gamme d'acquisition	50 à 2000 Da
Temps d'acquisition	30 minutes
Capillaire	3 kV
Cône d'échantillonnage	45 V
Cône d'extraction	4,0 V
Température de source	120°C
Température de désolvatation	600°C
Débit N_2 cône d'échantillonnage	20 L/h

Débit N_2 cône désolvatation	900 L/h
LockSpray™	Leucine Enképhaline (2 ng/µL) Intervalle d'infusion = 20 sec

6. Traitement des données LC-MS

6.1. Paramètres pour le prétraitement par XCMS

L'ensemble des paramètres XCMS utilisés pour traiter les données dans le cadre de notre stratégie histonomique est détaillé ci-dessous.

Etape	Méthode	Paramètre	Valeur
Détection	centWave	ppm	25
		snthr	10
		peakwidth	10 - 45
		mzdiff	- 0,005
		prefilter peaks	3
		prefilter intensity	500
Correction	Obiwarp	profStep	1
Alignement	density	bw	2
		mzwid	0,015
		minfrac	0,3
		minsamp	1

6.2. Normalisations

Les différentes étapes de normalisation permettent d'obtenir une distribution gaussienne de l'intensité des variables se rapprochant de la loi Normale. Leurs formules respectives sont décrites ci-dessous. Toutes ces étapes ont été réalisées sous l'environnement R à partir de la matrice *X* des variables. Toutes les valeurs

normalisées $\tilde{x}$ sont calculées à partir des valeurs brutes x pour chaque variable *i* présente dans l'échantillon *j*.

Normalisation par la médiane Md : $$\tilde{x}_{ij} = \frac{x_{ij}}{Md(x_j)}$$

Transformation logarithmique : $$\tilde{x}_{ij} = \log_{10}(x_{ij})$$

Centrage sur la moyenne : $$\tilde{x}_{ij} = x_{ij} - \left(\frac{1}{j} \times \sum_{i=1}^{j} x_{ij}\right)$$

Redimensionnement de Pareto : $$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\sigma_i}}$$

7. Validation et interprétation des résultats

7.1.Coefficient de variation

Pour chaque ion sélectionné lors des analyses statistiques multivariées, le coefficient de variation (CV) de l'intensité de l'ion *i* correspondant à travers tous les réplicats de l'échantillon QC a été calculé selon la formule suivante :

$$CV_i^{QC}(\%) = 100 \times \frac{\sigma_i^{QC}}{\mu_i^{QC}}$$ (σ = écart-type, μ = moyenne)

7.2.Test t de Welch

Le test t de Welch définit le t statistique par la formule suivante :

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

(1 et 2 correspondent aux deux populations étudiées, μ à la moyenne d'un échantillon, σ à son écart-type et N à sa taille).

7.3.Ratios d'abondance

Les logarithmes binaires des ratios des moyennes de chaque variable i sélectionnée dans les deux groupes d'échantillons 1 (témoins) et 2 (exposés) ont été calculés de la manière suivante :

$$\mathrm{FC}_i = \log_2\left(\frac{\mu_2^i}{\mu_1^i}\right)$$

7.4.Identification des variables validées

7.4.1. Paramètres de déconvolution MaxEnt1

Les paramètres MaxEnt1 dépendent principalement de la qualité du spectre et peuvent varier d'un spectre à l'autre.

Paramètre	Commentaire	Valeur
Gamme de masse	Gamme de masse du spectre déconvolué	5000 – 30000 Da
Résolution (Da/canal)	Résolution du spectre de masse déconvolué. Ne doit pas excéder la résolution réelle	0,50
Modèle	Modèle utilisé pour lisser les pics sur le spectre déconvolué	Gaussienne uniforme
Largeur à mi-hauteur	Largeur à mi-hauteur d'un pic mesuré sur le spectre de masse original	0,030
Ratio d'intensité minimum (%)	Limite gauche et droite de la hauteur relative des pics adjacents aux extrémités du spectre	33

7.4.2. Moteur de recherche TagIdent

Les paramètres utilisés pour l'identification par TagIdent (http://web.expasy.org/tagident) des protéines entières à partir des masses moléculaires moyennes observées sur les spectres déconvolués sont les suivants :

Tableau 29 : résumé des paramètres utilisés pour l'identification des protéines sur TagIdent.

Paramètre	Valeur
Taxonomie	Homo Sapiens
Gamme de pI	9 – 12
Masse moléculaire (Da)	Masse moyenne observée sur les spectres déconvolués
Erreur relative (%)	0,005

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
2. Waddington, C. H. *An introduction to modern genetics*. (The Macmillan company, 1939). at <http://catalog.hathitrust.org/Record/001492400>
3. Krude, T. & Keller, C. Chromatin assembly during S phase: contributions from histone deposition, DNA replication and the cell division cycle. *Cell. Mol. Life Sci.* **58,** 665–672 (2001).
4. Imhof, A. & Bonaldi, T. 'Chromatomics' the analysis of the chromatome. *Mol. Biosyst.* **1,** 112–116 (2005).
5. Comings, D. E. The structure and function of chromatin. *Adv. Hum. Genet.* **3,** 237–431 (1972).
6. Allis, C. D., Jenuwein, T., Reinberg, D. & Caparros, M.-L. *Epigenetics*. (Cold Spring Harbor Laboratory Press, New York, 2007).
7. Brodeur, J. & Toussaint, M. *Biologie moléculaire, Concepts, Techniques, Applications*. (Chenelière Education, 2007). at <http://www.ccdmd.qc.ca/catalogue/biologie-moleculaire>
8. Babu, A. & Verma, R. S. Chromosome structure: euchromatin and heterochromatin. *Int. Rev. Cytol.* **108,** 1–60 (1987).
9. Jones, P. A. & Taylor, S. M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20,** 85–93 (1980).
10. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).
11. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321,** 209–213 (1986).
12. Herman, J. G. & Baylin, S. B. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *New Engl. J. Med.* **349,** 2042–2054 (2003).
13. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324,** 929–930 (2009).
14. Ito, S. *et al*. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333,** 1300–1303 (2011).
15. Ficz, G. *et al*. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473,** 398–402 (2011).
16. He, Y.-F. *et al*. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333,** 1303–1307 (2011).
17. Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* **11,** 607–620 (2010).
18. Amaral, P. P., Dinger, M. E., Mercer, T. R. & Mattick, J. S. The eukaryotic genome as an RNA machine. *Science* **319,** 1787–1789 (2008).
19. Shrey, K., Suchit, A., Nishant, M. & Vibha, R. RNA interference: emerging diagnostics and therapeutics tool. *Biochem. Biophys. Res. Commun.* **386,** 273–277 (2009).

20. Lim, L. P. *et al*. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433,** 769–773 (2005).
21. Benetti, R. *et al*. A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. *Nat. Struct. Mol. Biol.* **15,** 998 (2008).
22. Bénard, J. & Douc-Rasy, S. Micro-RNA and oncogenesis. *Bulletin du Cancer* **92,** 757–762 (2005).
23. Gilbert, N. & Ramsahoye, B. The relationship between chromatin structure and transcriptional activity in mammalian genomes. *Brief. Funct. Genomic. Proteomic.* **4,** 129–142 (2005).
24. Kinner, A., Wu, W., Staudt, C. & Iliakis, G. H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. *Nucleic Acids Res.* **36,** 5678–5694 (2008).
25. Ramaswamy, A., Bahar, I. & Ioshikhes, I. Structural dynamics of nucleosome core particle: comparison with nucleosomes containing histone variants. *Proteins* **58,** 683–696 (2005).
26. Cerf, C. *et al*. Homo- and heteronuclear two-dimensional NMR studies of the globular domain of histone H1: full assignment, tertiary structure, and comparison with the globular domain of histone H5. *Biochemistry* **33,** 11079–11086 (1994).
27. Franklin, S. G. & Zweidler, A. Non-allelic variants of histones 2a, 2b and 3 in mammals. *Nature* **266,** 273–275 (1977).
28. Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The human and mouse replication-dependent histone genes. *Genomics* **80,** 487–498 (2002).
29. Wu, R. S., Tsai, S. & Bonner, W. M. Patterns of histone variant synthesis can distinguish G0 from G1 cells. *Cell* **31,** 367–374 (1982).
30. Dominski, Z. & Marzluff, W. F. Formation of the 3' end of histone mRNA. *Gene* **239,** 1–14 (1999).
31. Rasmussen, T. P. *et al*. Messenger RNAs encoding mouse histone macroH2A1 isoforms are expressed at similar levels in male and female cells and result from alternative splicing. *Nucleic Acids Res.* **27,** 3685–3689 (1999).
32. Bönisch, C. & Hake, S. B. Histone H2A variants in nucleosomes and chromatin: more or less stable? *Nucleic Acids Res.* **40,** 10719–10741 (2012).
33. Biterge, B. & Schneider, R. Histone variants: key players of chromatin. *Cell Tissue Res.* **356,** 457–466 (2014).
34. Pinto, D. M. S. & Flaus, A. Structure and function of histone H2AX. *Subcell. Biochem.* **50,** 55–78 (2010).
35. Fink, M., Imholz, D. & Thoma, F. Contribution of the serine 129 of histone H2A to chromatin structure. *Mol. Cell. Biol.* **27,** 3589–3600 (2007).
36. Thakar, A. *et al*. H2A.Z and H3.3 histone variants affect nucleosome structure: biochemical and biophysical studies. *Biochemistry* **48,** 10852–10857 (2009).
37. Barski, A. *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).

38. Zilberman, D., Coleman-Derr, D., Ballinger, T. & Henikoff, S. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456,** 125–129 (2008).

39. Xu, Y. *et al*. Histone H2A.Z controls a critical chromatin remodeling step required for DNA double-strand break repair. *Mol. Cell* **48,** 723–733 (2012).

40. Muthurajan, U. M., McBryant, S. J., Lu, X., Hansen, J. C. & Luger, K. The linker region of macroH2A promotes self-association of nucleosomal arrays. *J. Biol. Chem.* **286,** 23852–23864 (2011).

41. Doyen, C.-M. *et al*. Mechanism of polymerase II transcription repression by the histone variant macroH2A. *Mol. Cell. Biol.* **26,** 1156–1164 (2006).

42. Barrero, M. J. *et al*. Macrohistone variants preserve cell identity by preventing the gain of H3K4me2 during reprogramming to pluripotency. *Cell Rep.* **3,** 1005–1011 (2013).

43. Chadwick, B. P. & Willard, H. F. A novel chromatin protein, distantly related to histone H2A, is largely excluded from the inactive X chromosome. *J. Cell Biol.* **152,** 375–384 (2001).

44. Tolstorukov, M. Y. *et al*. Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Mol. Cell* **47,** 596–607 (2012).

45. Khare, S. P. *et al*. HIstome--a relational knowledgebase of human histone proteins and histone modifying enzymes. *Nucleic Acids Res.* **40,** D337–D342 (2012).

46. Churikov, D. *et al*. Novel human testis-specific histone H2B encoded by the interrupted gene on the X chromosome. *Genomics* **84,** 745–756 (2004).

47. Hake, S. B. & Allis, C. D. Histone H3 variants and their potential role in indexing mammalian genomes: the 'H3 barcode hypothesis'. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 6428–6435 (2006).

48. McKittrick, E., Gafken, P. R., Ahmad, K. & Henikoff, S. Histone H3.3 is enriched in covalent modifications associated with active chromatin. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 1525–1530 (2004).

49. Chen, P. *et al*. H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev.* **27,** 2109–2124 (2013).

50. Jin, C. *et al*. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.* **41,** 941–945 (2009).

51. Amor, D. J., Kalitsis, P., Sumer, H. & Choo, K. H. A. Building the centromere: from foundation proteins to 3D organization. *Trends Cell Biol.* **14,** 359–368 (2004).

52. Yang, J. W. *et al*. Human mini-chromosomes with minimal centromeres. *Hum. Mol. Genet.* **9,** 1891–1902 (2000).

53. Misteli, T., Gunjan, A., Hock, R., Bustin, M. & Brown, D. T. Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408,** 877–881 (2000).

54. Izzo, A., Kamieniarz, K. & Schneider, R. The histone H1 family: specific members, specific functions? *Biol. Chem.* **389,** 333–343 (2008).

55. Malik, H. S. & Henikoff, S. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* **10,** 882–891 (2003).

56. Yan, W., Ma, L., Burns, K. H. & Matzuk, M. M. HILS1 is a spermatid-specific linker histone H1-like protein implicated in chromatin remodeling during mammalian spermiogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **100,** 10546–10551 (2003).

57. Martianov, I. *et al.* Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 2808–2813 (2005).

58. Mizusawa, Y. *et al.* Expression of human oocyte-specific linker histone protein and its incorporation into sperm chromatin during fertilization. *Fertil. Steril.* **93,** 1134–1141 (2010).

59. Gkikopoulos, T. *et al.* A role for Snf2 related nucleosome spacing enzymes in genome-wide nucleosome organization. *Science* **333,** 1758–1760 (2011).

60. Hargreaves, D. C. & Crabtree, G. R. ATP-dependent chromatin remodeling: genetics, genomics and mechanisms. *Cell Res.* **21,** 396–420 (2011).

61. Muchardt, C. & Yaniv, M. When the SWI/SNF complex remodels...the cell cycle. *Oncogene* **20,** 3067–3075 (2001).

62. Ryme, J., Asp, P., Böhm, S., Cavellán, E. & Farrants, A.-K. O. Variations in the composition of mammalian SWI/SNF chromatin remodelling complexes. *J. Cell. Biochem.* **108,** 565–576 (2009).

63. Pandita, T. K. & Richardson, C. Chromatin remodeling finds its place in the DNA double-strand break response. *Nucleic Acids Res.* **37,** 1363–1377 (2009).

64. Deuring, R. *et al.* The ISWI chromatin-remodeling protein is required for gene expression and the maintenance of higher order chromatin structure in vivo. *Mol. Cell* **5,** 355–365 (2000).

65. Corona, D. F. V. *et al.* ISWI Regulates Higher-Order Chromatin Structure and Histone H1 Assembly In Vivo. *PLoS Biol.* **5,** (2007).

66. Konev, A. Y. *et al.* The CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. *Science* **317,** 1087–1090 (2007).

67. Pegoraro, G. & Misteli, T. The central role of chromatin maintenance in aging. *Aging (Albany NY)* **1,** 1017–1022 (2009).

68. Hurd, E. A., Poucher, H. K., Cheng, K., Raphael, Y. & Martin, D. M. The ATP-dependent chromatin remodeling enzyme CHD7 regulates pro-neural gene expression and neurogenesis in the inner ear. *Development* **137,** 3139–3150 (2010).

69. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).

70. Karch, K. R., DeNizio, J. E., Black, B. E. & Garcia, B. A. Identification and interrogation of combinatorial histone modifications. *Front. Genet.* **4,** 264 (2013).

71. Brunner, A. M., Tweedie-Cullen, R. Y. & Mansuy, I. M. Epigenetic modifications of the neuroproteome. *Proteomics* **12,** 2404–2420 (2012).

72. Ogryzko, V. V. Mammalian histone acetyltransferases and their complexes. *Cell. Mol. Life Sci.* **58,** 683–692 (2001).

73. Kingston, R. E. & Narlikar, G. J. ATP-dependent remodeling and acetylation as regulators of chromatin fluidity. *Genes Dev.* **13,** 2339–2352 (1999).

74. Bannister, A. J., Schneider, R. & Kouzarides, T. Histone Methylation: Dynamic or Static? *Cell* **109,** 801–806 (2002).

75. Lu, H.-R., Wang, X. & Wang, Y. A stronger DNA damage-induced G2 checkpoint due to over-activated CHK1 in the absence of PARP-1. *Cell Cycle* **5,** 2364–2370 (2006).

76. Thompson, P. R. & Fast, W. Histone citrullination by protein arginine deiminase: is arginine methylation a green light or a roadblock? *ACS Chem. Biol.* **1,** 433–441 (2006).

77. Shiio, Y. & Eisenman, R. N. Histone sumoylation is associated with transcriptional repression. *PNAS* **100,** 13225–13230 (2003).

78. Weake, V. M. & Workman, J. L. Histone Ubiquitination: Triggering Gene Activity. *Molecular Cell* **29,** 653–663 (2008).

79. Castro, P. H., Tavares, R. M., Bejarano, E. R. & Azevedo, H. SUMO, a heavyweight player in plant abiotic stress responses. *Cell. Mol. Life Sci.* **69,** 3269–3283 (2012).

80. Messner, S. & Hottiger, M. O. Histone ADP-ribosylation in DNA repair, replication and transcription. *Trends Cell Biol.* **21,** 534–542 (2011).

81. Kothapalli, N. *et al.* Biological functions of biotinylated histones. *J. Nutr. Biochem.* **16,** 446–448 (2005).

82. Peters, D. M., Griffin, J. B., Stanley, J. S., Beck, M. M. & Zempleni, J. Exposure to UV light causes increased biotinylation of histones in Jurkat cells. *Am. J. Physiol., Cell Physiol.* **283,** C878–884 (2002).

83. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146,** 1016–1028 (2011).

84. Tamkun, J. W. *et al.* brahma: a regulator of Drosophila homeotic genes structurally related to the yeast transcriptional activator SNF2/SWI2. *Cell* **68,** 561–572 (1992).

85. Owen, D. J. *et al.* The structural basis for the recognition of acetylated histone H4 by the bromodomain of histone acetyltransferase gcn5p. *EMBO J.* **19,** 6141–6149 (2000).

86. Sanchez, R. & Zhou, M.-M. The role of human bromodomains in chromatin biology and gene transcription. *Curr. Opin. Drug Discov. Devel.* **12,** 659–665 (2009).

87. Völkel, P. & Angrand, P.-O. The control of histone lysine methylation in epigenetic regulation. *Biochimie* **89,** 1–20 (2007).

88. Paro, R. & Hogness, D. S. The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **88,** 263–267 (1991).

89. Nielsen, A. L. *et al*. Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J.* **18,** 6385–6395 (1999).

90. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295,** 2080–2083 (2002).

91. Wang, Y., Jiang, F., Zhuo, Z., Wu, X.-H. & Wu, Y.-D. A Method for WD40 Repeat Detection and Secondary Structure Prediction. *PLoS One* **8,** (2013).

92. Huang, Y., Fang, J., Bedford, M. T., Zhang, Y. & Xu, R.-M. Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A. *Science* **312,** 748–751 (2006).

93. Wang, W. K. *et al*. Malignant Brain Tumor Repeats: A Three-Leaved Propeller Architecture with Ligand/Peptide Binding Pockets. *Structure* **11,** 775–789 (2003).

94. Schindler, U., Beckmann, H. & Cashmore, A. R. HAT3.1, a novel Arabidopsis homeodomain protein containing a conserved cysteine-rich region. *Plant J.* **4,** 137–150 (1993).

95. Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D. & Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Structural & Molecular Biology* **14,** 1025–1040 (2007).

96. Booker, G. W. *et al*. Structure of an SH2 domain of the p85 alpha subunit of phosphatidylinositol-3-OH kinase. *Nature* **358,** 684–687 (1992).

97. Rajewsky, M. F. Specificity of DNA damage in chemical carcinogenesis. *IARC Sci. Publ*. 41–54 (1980).

98. Lutz, W. K. Dose-response relationships in chemical carcinogenesis: from DNA adducts to tumor incidence. *Adv. Exp. Med. Biol*. **283,** 151–156 (1991).

99. Loeb, L. A. & Harris, C. C. Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res.* **68,** 6863–6872 (2008).

100. Pruss-Ustun, A. & Corvalan, C. Preventing disease through healthy environments: Towards an estimate of the environmental burden of disease. *World Health Organization* (2006). at <http://www.who.int/quantifying_ehimpacts/publications/preventingdisease/en/>

101. Barker, D. J. & Osmond, C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* **1,** 1077–1081 (1986).

102. Barker, D. J., Winter, P. D., Osmond, C., Margetts, B. & Simmonds, S. J. Weight in infancy and death from ischaemic heart disease. *Lancet* **2,** 577–580 (1989).

103. Barker, D. J. *et al*. Fetal nutrition and cardiovascular disease in adult life. *Lancet* **341,** 938–941 (1993).

104. Hanson, M., Godfrey, K. M., Lillycrop, K. A., Burdge, G. C. & Gluckman, P. D. Developmental plasticity and developmental origins of non-communicable

disease: theoretical considerations and epigenetic mechanisms. *Prog. Biophys. Mol. Biol.* **106,** 272–280 (2011).

105. Evain-Brion, D. & Malassiné, A. *Le placenta humain*. (Tec & Doc Lavoisier, 2010).
106. Alsat, E. & Evain-Brion, D. Le placenta humain : neuf mois d'une intense activité encore méconnue. *Médecine thérapeutique / Pédiatrie* **1,** 509–16 (1999).
107. Pasanen, M. The expression and regulation of drug metabolism in human placenta. *Adv. Drug Deliv. Rev.* **38,** 81–97 (1999).
108. Hakkola, J. *et al.* Detection of cytochrome P450 gene expression in human placenta in first trimester of pregnancy. *Biochem. Pharmacol.* **52,** 379–383 (1996).
109. Hakkola, J. *et al.* Expression of xenobiotic-metabolizing cytochrome P450 forms in human full-term placenta. *Biochem. Pharmacol.* **51,** 403–411 (1996).
110. Collier, A. C., Tingle, M. D., Paxton, J. W., Mitchell, M. D. & Keelan, J. A. Metabolizing enzyme localization and activities in the first trimester human placenta: the effect of maternal and gestational age, smoking and alcohol consumption. *Hum. Reprod.* **17,** 2564–2572 (2002).
111. Myatt, L. Placental adaptive responses and fetal programming. *J. Physiol. (Lond.)* **572,** 25–30 (2006).
112. Sandovici, I., Hoelle, K., Angiolini, E. & Constância, M. Placental adaptations to the maternal-fetal environment: implications for fetal growth and developmental programming. *Reprod. Biomed. Online* **25,** 68–89 (2012).
113. Kacem, S. & Feil, R. Chromatin mechanisms in genomic imprinting. *Mamm. Genome* **20,** 544–556 (2009).
114. Gimelbrant, A., Hutchinson, J. N., Thompson, B. R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318,** 1136–1140 (2007).
115. Nelissen, E. C. M., van Montfoort, A. P. A., Dumoulin, J. C. M. & Evers, J. L. H. Epigenetics and the placenta. *Hum. Reprod. Update* **17,** 397–417 (2011).
116. Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.* **13,** 97–109 (2011).
117. Szyf, M. The Dynamic Epigenome and its Implications in Toxicology. *Toxicol. Sci.* **100,** 7–23 (2007).
118. Kanthasamy, A. *et al.* Emerging neurotoxic mechanisms in environmental factors-induced neurodegeneration. *Neurotoxicology* **33,** 833–837 (2012).
119. Baccarelli, A. & Bollati, V. Epigenetics and environmental chemicals. *Curr. Opin. Pediatr.* **21,** 243–251 (2009).
120. Celander, M. C. Cocktail effects on biomarker responses in fish. *Aquat. Toxicol.* **105,** 72–77 (2011).
121. Jang, H. & Serra, C. Nutrition, epigenetics, and diseases. *Clin. Nutr. Res.* **3,** 1–8 (2014).
122. Broday, L. *et al.* Nickel compounds are novel inhibitors of histone H4 acetylation. *Cancer Res.* **60,** 238–241 (2000).

123. Zhang, Q. *et al*. Inhibition and reversal of nickel-induced transformation by the histone deacetylase inhibitor trichostatin A. *Toxicol. Appl. Pharmacol.* **192,** 201–211 (2003).

124. Ke, Q., Davidson, T., Chen, H., Kluz, T. & Costa, M. Alterations of histone modifications and transgene silencing by nickel chloride. *Carcinogenesis* **27,** 1481–1488 (2006).

125. Arita, A. *et al*. Global levels of histone modifications in peripheral blood mononuclear cells of subjects with exposure to nickel. *Environ. Health Perspect.* **120,** 198–203 (2012).

126. Hubaux, R. *et al*. Molecular features in arsenic-induced lung tumors. *Mol. Cancer* **12,** 20 (2013).

127. Zhou, X., Li, Q., Arita, A., Sun, H. & Costa, M. Effects of nickel, chromate, and arsenite on histone 3 lysine methylation. *Toxicology and Applied Pharmacology* **236,** 78–84 (2009).

128. Jensen, T. J., Novak, P., Eblin, K. E., Gandolfi, A. J. & Futscher, B. W. Epigenetic remodeling during arsenical-induced malignant transformation. *Carcinogenesis* **29,** 1500–1508 (2008).

129. Jensen, T. J. *et al*. Epigenetic mediated transcriptional activation of WNT5A participates in arsenical-associated malignant transformation. *Toxicol. Appl. Pharmacol.* **235,** 39–46 (2009).

130. Chu, F. *et al*. Quantitative mass spectrometry reveals the epigenome as a target of arsenic. *Chem. Biol. Interact.* **192,** 113–117 (2011).

131. Somji, S. *et al*. Differences in the epigenetic regulation of MT-3 gene expression between parental and Cd+2 or As+3 transformed human urothelial cells. *Cancer Cell Int.* **11,** 2 (2011).

132. Schnekenburger, M., Talaska, G. & Puga, A. Chromium cross-links histone deacetylase 1-DNA methyltransferase 1 complexes to chromatin, inhibiting histone-remodeling marks critical for transcriptional activation. *Mol. Cell. Biol.* **27,** 7089–7101 (2007).

133. Sun, H., Zhou, X., Chen, H., Li, Q. & Costa, M. Modulation of histone methylation and MLH1 gene silencing by hexavalent chromium. *Toxicol. Appl. Pharmacol.* **237,** 258–266 (2009).

134. Zakhari, S. Alcohol metabolism and epigenetics changes. *Alcohol Res.* **35,** 6–16 (2013).

135. D'Addario, C. *et al*. Ethanol and acetaldehyde exposure induces specific epigenetic modifications in the prodynorphin gene promoter in a human neuroblastoma cell line. *FASEB J.* **25,** 1069–1075 (2011).

136. Pal-Bhadra, M. *et al*. Distinct methylation patterns in histone H3 at Lys-4 and Lys-9 correlate with up- & down-regulation of genes by ethanol in hepatocytes. *Life Sci.* **81,** 979–987 (2007).

137. Li, Q., Ke, Q. & Costa, M. Alterations of histone modifications by cobalt compounds. *Carcinogenesis* **30,** 1243–1251 (2009).

138. Rogge, G. A. & Wood, M. A. The role of histone acetylation in cocaine-induced neural plasticity and behavior. *Neuropsychopharmacology* **38,** 94–110 (2013).

139. Maze, I. *et al*. Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 3035–3040 (2011).

140. Brami-Cherrier, K. *et al*. Parsing molecular and behavioral effects of cocaine in mitogen- and stress-activated protein kinase-1-deficient mice. *J. Neurosci.* **25,** 11444–11454 (2005).

141. Yang, X. *et al*. Histone modifications are associated with Δ9-tetrahydrocannabinol-mediated alterations in antigen-specific T cell responses. *J. Biol. Chem.* **289,** 18707–18718 (2014).

142. Collotta, M., Bertazzi, P. A. & Bollati, V. Epigenetics and pesticides. *Toxicology* **307,** 35–41 (2013).

143. Song, C., Kanthasamy, A., Anantharam, V., Sun, F. & Kanthasamy, A. G. Environmental neurotoxic pesticide increases histone acetylation to promote apoptosis in dopaminergic neuronal cells: relevance to epigenetic mechanisms of neurodegeneration. *Mol. Pharmacol.* **77,** 621–632 (2010).

144. Song, C., Kanthasamy, A., Jin, H., Anantharam, V. & Kanthasamy, A. G. Paraquat induces epigenetic changes by promoting histone acetylation in cell culture models of dopaminergic degeneration. *Neurotoxicology* **32,** 586–595 (2011).

145. Maloney, B., Sambamurti, K., Zawia, N. & Lahiri, D. K. Applying epigenetics to Alzheimer's disease via the latent early-life associated regulation (LEARn) model. *Curr. Alzheimer Res.* **9,** 589–599 (2012).

146. Lahiri, D. K. & Maloney, B. The 'LEARn' (latent early-life associated regulation) model: an epigenetic pathway linking metabolic and cognitive disorders. *J. Alzheimers Dis.* **30 Suppl 2,** S15–30 (2012).

147. Guillén, M. D., Sopelana, P. & Partearroyo, M. A. Food as a source of polycyclic aromatic carcinogens. *Rev. Environ. Health* **12,** 133–146 (1997).

148. Sadikovic, B., Andrews, J., Carter, D., Robinson, J. & Rodenhiser, D. I. Genome-wide H3K9 histone acetylation profiles are altered in benzopyrene-treated MCF7 breast cancer cells. *J. Biol. Chem.* **283,** 4051–4060 (2008).

149. Dik, S., Scheepers, P. T. J. & Godderis, L. Effects of environmental stressors on histone modifications and their relevance to carcinogenesis: A systematic review. *Crit. Rev. Toxicol.* **42,** 491–500 (2012).

150. Magi, B. & Liberatori, S. Immunoblotting techniques. *Methods Mol. Biol.* **295,** 227–254 (2005).

151. Fuchs, S. M., Krajewski, K., Baker, R. W., Miller, V. L. & Strahl, B. D. Influence of combinatorial histone modifications on antibody and effector protein recognition. *Curr. Biol.* **21,** 53–58 (2011).

152. Wilkins, M. R. *et al*. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N.Y.)* **14,** 61–65 (1996).

153. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).

154. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301 (1988).

155. Stults, J. T. Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). *Curr. Opin. Struct. Biol.* **5,** 691–698 (1995).

156. Börnsen, K. O. Influence of salts, buffers, detergents, solvents, and matrices on MALDI-MS protein analysis in complex mixtures. *Methods Mol. Biol.* **146,** 387–404 (2000).

157. De Hoffmann, E. & Stroobant, V. *Spectrométrie de masse. Cours et exercices corrigés - 3e édition.* (Dunod, 2005). at <http://www.decitre.fr/livres/spectrometrie-de-masse-9782100494491.html>

158. Clegg, G. A. & Dole, M. Molecular beams of macroions. 3. Zein and polyvinylpyrrolidone. *Biopolymers* **10,** 821–826 (1971).

159. Whitehouse, C. M., Dreyer, R. N., Yamashita, M. & Fenn, J. B. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679 (1985).

160. Mora, J. F. *et al.* Electrochemical processes in electrospray ionization mass spectrometry. *J. Mass Spectrom.* **35,** 939–952 (2000).

161. Cech, N. B. & Enke, C. G. Effect of affinity for droplet surfaces on the fraction of analyte molecules charged during electrospray droplet fission. *Anal. Chem.* **73,** 4632–4639 (2001).

162. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J. Mass Spectrom.* **35,** 804–817 (2000).

163. Emmett, M. R. & Caprioli, R. M. Micro-electrospray mass spectrometry: Ultra-high-sensitivity analysis of peptides and proteins. *J. Am. Soc. Mass Spectrom.* **5,** 605–613 (1994).

164. Körner, R., Wilm, M., Morand, K., Schubert, M. & Mann, M. Nano electrospray combined with a quadrupole ion trap for the analysis of peptides and protein digests. *J. Am. Soc. Mass Spectrom.* **7,** 150–156 (1996).

165. Paul, W. & Steinwedel, H. Ein neues Massenspektrometer ohne Magnetfeld. *Zeitschrift Naturforschung Teil A* **8,** 448 (1953).

166. Wiley, W. C. & McLaren, I. H. Time-of-Flight Mass Spectrometer with Improved Resolution. *Rev.Sci.Instrum.* **26,** 1150–1157 (1955).

167. Cornish, T. J. & Cotter, R. J. A curved field reflectron time-of-flight mass spectrometer for the simultaneous focusing of metastable product ions. *Rapid Commun. Mass Spectrom.* **8,** 781–785 (1994).

168. Guilhaus, M., Selby, D. & Mlynski, V. Orthogonal acceleration time-of-flight mass spectrometry. *Mass Spectrom. Rev.* **19,** 65–107 (2000).

169. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).

170. Neuhoff, V., Arold, N., Taube, D. & Ehrhardt, W. Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250. *Electrophoresis* **9,** 255–262 (1988).

171. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858 (1996).

172. Zweidler, A. Resolution of histones by polyacrylamide gel electrophoresis in presence of nonionic detergents. *Methods Cell Biol.* **17,** 223–233 (1978).

173. Waterborg, J. H., Winicov, I. & Harrington, R. E. Histone variants and acetylated species from the alfalfa plant Medicago sativa. *Arch. Biochem. Biophys.* **256,** 167–178 (1987).

174. Lindner, H., Sarg, B., Meraner, C. & Helliger, W. Separation of acetylated core histones by hydrophilic-interaction liquid chromatography. *J. Chromatogr. A* **743,** 137–144 (1996).

175. Plumb, R. *et al.* Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **18,** 2331–2337 (2004).

176. Contrepois, K., Ezan, E., Mann, C. & Fenaille, F. Ultra-high performance liquid chromatography-mass spectrometry for the fast profiling of histone post-translational modifications. *J. Proteome Res.* **9,** 5501–5509 (2010).

177. Nordström, A., O'Maille, G., Qin, C. & Siuzdak, G. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. *Anal. Chem.* **78,** 3289–3295 (2006).

178. Mann, M., Højrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345 (1993).

179. Yates, J. R., Speicher, S., Griffin, P. R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408 (1993).

180. Gattiker, A., Bienvenut, W. V., Bairoch, A. & Gasteiger, E. FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics* **2,** 1435–1444 (2002).

181. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11,** 601 (1984).

182. Biemann, K. Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. *Meth. Enzymol.* **193,** 455–479 (1990).

183. Biemann, K. & Scoble, H. A. Characterization by tandem mass spectrometry of structural modifications in proteins. *Science* **237,** 992–998 (1987).

184. McLafferty, F. W. *et al.* Two-dimensional mass spectrometry of biomolecules at the subfemtomole level. *Curr. Opin. Chem. Biol.* **2,** 571–578 (1998).

185. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 9528–9533 (2004).

186. McLuckey, S. A. & Huang, T.-Y. Ion/Ion Reactions: New Chemistry for Analytical MS. *Anal. Chem.* **81,** 8669–8676 (2009).

187. McAlister, G. C., Phanstiel, D., Good, D. M., Berggren, W. T. & Coon, J. J. Implementation of electron-transfer dissociation on a hybrid linear ion trap-orbitrap mass spectrometer. *Anal. Chem.* **79,** 3525–3534 (2007).

188. Garcia, B. A. *et al.* Chemical derivatization of histones for facilitated analysis by mass spectrometry. *Nat. Protoc.* **2,** 933–938 (2007).

189. McLafferty, F. W. *et al.* Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *FEBS J.* **274,** 6256–6268 (2007).

190. Banks, G. C., Deterding, L. J., Tomer, K. B. & Archer, T. K. Hormone-mediated dephosphorylation of specific histone H1 isoforms. *J. Biol. Chem.* **276,** 36467–36473 (2001).

191. Boyne, M. T., 2nd, Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Precise characterization of human histones in the H2A gene family by top down mass spectrometry. *J. Proteome Res.* **5,** 248–253 (2006).

192. Siuti, N., Roth, M. J., Mizzen, C. A., Kelleher, N. L. & Pesavento, J. J. Gene-specific characterization of human histone H2B by electron capture dissociation. *J. Proteome Res.* **5,** 233–239 (2006).

193. Thomas, C. E., Kelleher, N. L. & Mizzen, C. A. Mass spectrometric characterization of human histone H3: a bird's eye view. *J. Proteome Res.* **5,** 240–247 (2006).

194. Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Quantitative analysis of modified proteins and their positional isomers by tandem mass spectrometry: human histone H4. *Anal. Chem.* **78,** 4271–4280 (2006).

195. Tian, Z. *et al.* Two-dimensional liquid chromatography system for online top-down mass spectrometry. *Proteomics* **10,** 3610–3620 (2010).

196. Eliuk, S. M., Maltby, D., Panning, B. & Burlingame, A. L. High resolution electron transfer dissociation studies of unfractionated intact histones from murine embryonic stem cells using on-line capillary LC separation: determination of abundant histone isoforms and post-translational modifications. *Mol. Cell Proteomics* **9,** 824–837 (2010).

197. Zee, B. M., Young, N. L. & Garcia, B. A. Quantitative proteomic approaches to studying histone modifications. *Curr. Chem. Genomics* **5,** 106–114 (2011).

198. Molden, R. C. & Garcia, B. A. Middle-Down and Top-Down Mass Spectrometric Analysis of Co-occurring Histone Modifications. *Curr. Protoc. Protein Sci.* **77,** 23.7.1–23.7.28 (2014).

199. Taverna, S. D. *et al.* Long-distance combinatorial linkage between methylation and acetylation on histone H3 N termini. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 2086–2091 (2007).

200. Young, N. L. *et al.* High throughput characterization of combinatorial histone codes. *Mol. Cell Proteomics* **8,** 2266–2284 (2009).

201. Yang, L. *et al.* Unambiguous determination of isobaric histone modifications by reversed-phase retention time and high-mass accuracy. *Anal. Biochem.* **396,** 13–22 (2010).

202. Garcia, B. A. What does the future hold for Top Down mass spectrometry? *J. Am. Soc. Mass Spectrom.* **21,** 193–202 (2010).

203. Sidoli, S. *et al*. Middle-down hybrid chromatography/tandem mass spectrometry workflow for characterization of combinatorial post-translational modifications in histones. *Proteomics* **14,** 2200–2211 (2014).

204. Sidoli, S., Cheng, L. & Jensen, O. N. Proteomics in chromatin biology and epigenetics: Elucidation of post-translational modifications of histone proteins by mass spectrometry. *J Proteomics* **75,** 3419–3433 (2012).

205. Ong, S.-E. *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics* **1,** 376–386 (2002).

206. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404,** 939–965 (2012).

207. Arnaudo, A. M., Molden, R. C. & Garcia, B. A. Revealing histone variant induced changes via quantitative proteomics. *Crit. Rev. Biochem. Mol. Biol.* **46,** 284–294 (2011).

208. Eberl, H. C., Mann, M. & Vermeulen, M. Quantitative proteomics for epigenetics. *Chembiochem* **12,** 224–234 (2011).

209. Smith, C. M. *et al*. Mass spectrometric quantification of acetylation at specific lysines within the amino-terminal tail of histone H4. *Anal. Biochem.* **316,** 23–33 (2003).

210. Plazas-Mayorca, M. D. *et al*. One-pot shotgun quantitative mass spectrometry characterization of histones. *J. Proteome Res.* **8,** 5367–5374 (2009).

211. Bonenfant, D. *et al*. Analysis of dynamic changes in post-translational modifications of human histones during cell cycle by mass spectrometry. *Mol. Cell Proteomics* **6,** 1917–1932 (2007).

212. Zee, B. M., Levin, R. S., Dimaggio, P. A. & Garcia, B. A. Global turnover of histone post-translational modifications and variants in human cells. *Epigenetics Chromatin* **3,** 22 (2010).

213. Montes de Oca, R., Shoemaker, C. J., Gucek, M., Cole, R. N. & Wilson, K. L. Barrier-to-autointegration factor proteome reveals chromatin-regulatory partners. *PLoS ONE* **4,** e7050 (2009).

214. Ross, P. L. *et al*. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics* **3,** 1154–1169 (2004).

215. Pattillo, R. A. & Gey, G. O. The establishment of a cell line of human hormone-synthesizing trophoblastic cells in vitro. *Cancer Res.* **28,** 1231–6 (1968).

216. Bode, C. J. *et al*. In vitro models for studying trophoblast transcellular transport. *Methods Mol. Med.* **122,** 225–239 (2006).

217. Avery, M. L., Meek, C. E. & Audus, K. L. The presence of inducible cytochrome P450 types 1A1 and 1A2 in the BeWo cell line. *Placenta* **24,** 45–52 (2003).

218. Friedman, S. J. & Skehan, P. Morphological differentiation of human choriocarcinoma cells induced by methotrexate. *Cancer Res.* **39,** 1960–1967 (1979).

219. Shechter, D., Dormann, H. L., Allis, C. D. & Hake, S. B. Extraction, purification and analysis of histones. *Nat. Protoc.* **2,** 1445–1457 (2007).

220. Murray, K. The acid extraction of histones from calf thymus deoxyribonucleoprotein. *J. Mol. Biol.* **15,** 409–419 (1966).

221. Chen, C. C., Smith, D. L., Bruegger, B. B., Halpern, R. M. & Smith, R. A. Occurrence and distribution of acid-labile histone phosphates in regenerating rat liver. *Biochemistry* **13,** 3785–3789 (1974).

222. Matthews, H. R. & Huebner, V. D. Nuclear protein kinases. *Mol. Cell. Biochem.* **59,** 81–99 (1984).

223. Von Holt, C. *et al.* Isolation and characterization of histones. *Meth. Enzymol.* **170,** 431–523 (1989).

224. Sivaraman, T., Kumar, T. K., Jayaraman, G. & Yu, C. The mechanism of 2,2,2-trichloroacetic acid-induced protein precipitation. *J. Protein Chem.* **16,** 291–297 (1997).

225. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).

226. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).

227. Smith, P. K. *et al.* Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150,** 76–85 (1985).

228. Everley, R. A. & Croley, T. R. Ultra-performance liquid chromatography/mass spectrometry of intact proteins. *J. Chromatogr. A* **1192,** 239–247 (2008).

229. Eschelbach, J. W. & Jorgenson, J. W. Improved protein recovery in reversed-phase liquid chromatography by the use of ultrahigh pressures. *Anal. Chem.* **78,** 1697–1706 (2006).

230. Rehder, D. S., Dillon, T. M., Pipes, G. D. & Bondarenko, P. V. Reversed-phase liquid chromatography/mass spectrometry analysis of reduced monoclonal antibodies in pharmaceutics. *J. Chromatogr. A* **1102,** 164–175 (2006).

231. Rusconi, F. *Manuel de spectrométrie de masse à l'usage des biochimistes.* (Lavoisier, 2011).

232. Page, J. S., Kelly, R. T., Tang, K. & Smith, R. D. Ionization and Transmission Efficiency in an Electrospray Ionization–Mass Spectrometry Interface. *J. Am. Soc. Mass Spectrom.* **18,** 1582–1590 (2007).

233. Montgomery, D. C. *Design and Analysis of Experiments.* (John Wiley & Sons, 2008).

234. Stanstrup, J., Gerlich, M., Dragsted, L. O. & Neumann, S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. *Anal. Bioanal. Chem.* **405,** 5037–48 (2013).

235. Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9,** 504 (2008).

236. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78,** 779–87 (2006).

237. Prince, J. T. & Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **78,** 6140–6152 (2006).

238. Katajamaa, M. & Oresic, M. Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* **1158,** 318–328 (2007).

239. Sauve, A. C. & Speed, T. P. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. in *In Procedings Gensips* (2004).

240. De Brauwere, A., Pintelon, R., De Ridder, F., Schoukens, J. & Baeyens, W. Estimation of heteroscedastic measurement noise variances. *Chemometr. Intell. Lab. Syst.* **86,** 130–138 (2007).

241. Kvalheim, O. M., Brakstad, F. & Liang, Y. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Anal. Chem.* **66,** 43–51 (1994).

242. Bro, R. & Smilde, A. K. Centering and scaling in component analysis. *J. Chemometrics* **17,** 16–33 (2003).

243. Eriksson, L., Johansson, E. & Kettapeh-Wold, S. *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*. (Umetrics, 1999).

244. Van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7,** 142 (2006).

245. Wold, S. Chemometrics; what do we mean with it, and what do we want from it? *Chemometr. Intell. Lab. Syst.* **30,** 109–115 (1995).

246. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58,** 236–244 (1963).

247. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24,** 417–441 (1933).

248. Wold, S. Cross-validatory estimation of the number of components in factor and principal components analysis. *Technometrics* **20,** 198–200 (1978).

249. Wold, S., Martens, H. & Wold, H. in *Matrix Pencils* (eds. Kågström, B. & Ruhe, A.) 286–293 (Springer Berlin Heidelberg, 1983). at <http://link.springer.com/chapter/10.1007/BFb0062108>

250. Ståhle, L. & Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1,** 185–196 (1987).

251. Eriksson, L., Trygg, J., Wold, S. CV-ANOVA for significance testing of PLS and OPLS models. *J. Chemom.* **22,** 594–600 (2008).

252. Trygg, J. & Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **16,** 119–128 (2002).

253. Wiklund, S. *et al*. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal. Chem.* **80,** 115–122 (2008).

254. Kirwan, J. A., Broadhurst, D. I., Davidson, R. L. & Viant, M. R. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Anal. Bioanal. Chem.* **405,** 5147–5157 (2013).

255. Welch, B. L. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika* **34,** 28–35 (1947).

256. Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. *Stat. Med.* **9,** 811–818 (1990).

257. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Meth.* **57,** 289–300 (1995).

258. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106,** 620–630 (1957).

259. Gasteiger, E. *et al*. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005). at <http://link.springer.com/protocol/10.1385/1-59259-890-0%3A571>

260. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191–D198 (2013).

261. Marchion, D. & Münster, P. Development of histone deacetylase inhibitors for cancer treatment. *Expert Rev. Anticancer Ther.* **7,** 583–598 (2007).

262. Davie, J. R. Inhibition of histone deacetylase activity by butyrate. *J. Nutr.* **133,** 2485S–2493S (2003).

263. Takai, N. & Narahara, H. Epigenetic Therapy in Human Choriocarcinoma. *Cancers* **2,** 1683–1688 (2010).

264. Bolden, J. E., Peart, M. J. & Johnstone, R. W. Anticancer activities of histone deacetylase inhibitors. *Nat. Rev. Drug Discov.* **5,** 769–784 (2006).

265. Carafa, V., Miceli, M., Altucci, L. & Nebbioso, A. Histone deacetylase inhibitors: a patent review (2009 - 2011). *Expert Opin. Ther. Pat.* **23,** 1–17 (2013).

266. West, A. C. & Johnstone, R. W. New and emerging HDAC inhibitors for cancer treatment. *J. Clin. Invest.* **124,** 30–39 (2014).

267. Chaturvedi, P. & Tyagi, S. C. Epigenetic mechanisms underlying cardiac degeneration and regeneration. *Int. J. Cardiol.* (2014). doi:10.1016/j.ijcard.2014.02.008

268. Konsoula, Z. & Barile, F. A. Epigenetic histone acetylation and deacetylation mechanisms in experimental models of neurodegenerative disorders. *J. Pharmacol. Toxicol. Methods* **66,** 215–220 (2012).

269. Harrison, I. F. & Dexter, D. T. Epigenetic targeting of histone deacetylase: therapeutic potential in Parkinson's disease? *Pharmacol. Ther.* **140,** 34–52 (2013).

270. Adwan, L. & Zawia, N. H. Epigenetics: a novel therapeutic approach for the treatment of Alzheimer's disease. *Pharmacol. Ther.* **139,** 41–50 (2013).

271. Candido, E. P. M., Reeves, R. & Davie, J. R. Sodium butyrate inhibits histone deacetylation in cultured cells. *Cell* **14,** 105–113 (1978).

272. Kruh, J. Effects of sodium butyrate, a new pharmacological agent, on cells in culture. *Mol. Cell. Biochem.* **42,** 65–82 (1981).

273. Bose, P., Dai, Y. & Grant, S. Histone deacetylase inhibitor (HDACI) mechanisms of action: Emerging insights. *Pharmacology & Therapeutics* doi:10.1016/j.pharmthera.2014.04.004

274. Wheelock, Å. M. & Wheelock, C. E. Trials and tribulations of 'omics data analysis: assessing quality of SIMCA-based multivariate models using examples from pulmonary medicine. *Mol. Biosyst.* **9,** 2589–2596 (2013).

275. Jiang, T., Zhou, X., Taghizadeh, K., Dong, M. & Dedon, P. C. N-formylation of lysine in histone proteins as a secondary modification arising from oxidative DNA damage. *Proc. Natl. Acad. Sci. U.S.A.* **104,** 60–65 (2007).

276. Pesavento, J. J., Yang, H., Kelleher, N. L. & Mizzen, C. A. Certain and progressive methylation of histone H4 at lysine 20 during the cell cycle. *Mol. Cell. Biol.* **28,** 468–486 (2008).

277. Zedeck, M. S. Polycyclic aromatic hydrocarbons: a review. *J. Environ. Pathol. Toxicol.* **3,** 537–567 (1980).

278. Nebert, D. W., Dalton, T. P., Okey, A. B. & Gonzalez, F. J. Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J. Biol. Chem.* **279,** 23847–23850 (2004).

279. Aimová, D. *et al.* Ellipticine and benzo(a)pyrene increase their own metabolic activation via modulation of expression and enzymatic activity of cytochromes P450 1A1 and 1A2. *Interdiscip. Toxicol.* **1,** 160–168 (2008).

280. Hockley, S. L. *et al.* AHR- and DNA-damage-mediated gene expression responses induced by benzo(a)pyrene in human cell lines. *Chem. Res. Toxicol.* **20,** 1797–1810 (2007).

281. Tung, E. W. Y., Philbrook, N. A., Belanger, C. L., Ansari, S. & Winn, L. M. Benzo[a]pyrene increases DNA double strand break repair in vitro and in vivo: a possible mechanism for benzo[a]pyrene-induced toxicity. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **760,** 64–69 (2014).

282. Jiang, Y., Wang, K., Fang, R. & Zheng, J. Expression of aryl hydrocarbon receptor in human placentas and fetal tissues. *J. Histochem. Cytochem.* **58,** 679–685 (2010).

283. Stejskalova, L. & Pavek, P. The function of cytochrome P450 1A1 enzyme (CYP1A1) and aryl hydrocarbon receptor (AhR) in the placenta. *Curr. Pharm. Biotechnol.* **12,** 715–730 (2011).

284. Rennie, M. Y. *et al.* Vessel tortuousity and reduced vascularization in the fetoplacental arterial tree after maternal exposure to polycyclic aromatic hydrocarbons. *Am. J. Physiol. Heart Circ. Physiol.* **300,** H675–684 (2011).

285. Malassiné, A., Frendo, J. L. & Evain-Brion, D. A comparison of placental development and endocrine functions between the human and mouse model. *Hum. Reprod. Update* **9,** 531–539 (2003).

286. Zhang, L. & Shiverick, K. T. Differential effects of 2,3,7,8-tetrachlorodibenzo-p-dioxin and benzo(a)pyrene on proliferation and growth factor gene expression in human choriocarcinoma BeWo cells. *Placenta* **19, Supplement 1,** 177–191 (1998).

287. Ovesen, J. L., Schnekenburger, M. & Puga, A. Aryl Hydrocarbon Receptor Ligands of Widely Different Toxic Equivalency Factors Induce Similar Histone Marks in Target Gene Chromatin. *Toxicol. Sci.* **121,** 123–131 (2011).

288. Lijinsky, W. The formation and occurrence of polynuclear aromatic hydrocarbons associated with food. *Mutat. Res.* **259,** 251–261 (1991).

289. Lodovici, M., Akpan, V., Evangelisti, C. & Dolara, P. Sidestream tobacco smoke as the main predictor of exposure to polycyclic aromatic hydrocarbons. *J. Appl. Toxicol.* **24,** 277–281 (2004).

290. Le Vee, M., Kolasa, E., Jouan, E., Collet, N. & Fardel, O. Differentiation of human placental BeWo cells by the environmental contaminant benzo(a)pyrene. *Chem. Biol. Interact.* **210,** 1–11 (2014).

291. Kjeldahl, K. & Bro, R. Some common misunderstandings in chemometrics. *J. Chemom.* **24,** 558–564 (2010).

292. Draker, R. *et al.* A combination of H2A.Z and H4 acetylation recruits Brd2 to chromatin during transcriptional activation. *PLoS Genet.* **8,** e1003047 (2012).

293. Sinha, A., Faller, D. V. & Denis, G. V. Bromodomain analysis of Brd2-dependent transcriptional activation of cyclin A. *Biochem. J.* **387,** 257–269 (2005).

294. Valdés-Mora, F. *et al.* Acetylation of H2A.Z is a key epigenetic modification associated with gene deregulation and epigenetic remodeling in cancer. *Genome Res.* **22,** 307–321 (2012).

295. Talbert, P. B. & Henikoff, S. Environmental responses mediated by histone variants. *Trends Cell Biol.* (2014). doi:10.1016/j.tcb.2014.07.006

296. Delacour, H., Servonnet, A., Perrot, A., Vigezzi, J. F. & Ramirez, J. M. [ROC (receiver operating characteristics) curve: principles and application in biology]. *Ann. Biol. Clin. (Paris)* **63,** 145–154 (2005).

297. Swets, J. A. Measuring the accuracy of diagnostic systems. *Science* **240,** 1285–1293 (1988).

ANNEXES

Liste des communications :

- 20th International Mass spectrometry Conference, Geneva August 24-29, 2014 **(Poster)**

Epigenetic effects of Benzo[a]pyrene on placental histones: a new global MS-based profiling approach

R. Bilgraer, S. Gillet, S. Gil, D. Evain-Brion, O. Laprévote

- 62nd ASMS Conference on Mass Spectrometry and Allied topics, Baltimore June 15-19, 2014 **(Poster)**

Dynamic changes in histone post-translational modifications: a new global MS-based profiling approach

R. Bilgraer, S. Gillet, S. Gil, D. Evain-Brion, O. Laprévote

- 1ère Journée Scientifique du Centre de recherche Pharmaceutique de Paris, Paris 12 Mai 2014 **(Oral)**

Déchiffrer le code histone: une démarche agnostique pour révéler des marqueurs d'exposition aux xénobiotiques

R. Bilgraer, S. Gillet, S. Gil, D. Evain-Brion, O. Laprévote

- Journées Scientifiques de l'Ecole doctorale MTCI, Paris 29-30 Avril 2014 **(Oral)**

Approche chimiométrique pour la recherche de biomarqueurs histoniques d'exposition placentaire

R. Bilgraer

- 4e Journée Scientifique de l'IMTCE, Paris 31 Mai 2013 **(Poster)**

Approche globale en protéomique différentielle: application à l'exploration du code histone

R. Bilgraer, S. Gillet, O. Laprévote

- 18e Rencontres du Club jeune de la SFSM, Izeste 8-12 Avril 2013 **(Oral)**

Caractérisation de modifications post-traductionnelles d'histones par spectrométrie de masse: application à la toxicité placentaire du benzo[a]pyrène

R. Bilgraer, S. Gillet, O. Laprévote

- 29e Journées de la SFSM, Orléans 17-20 Septembre 2012 **(Poster)**

Toxicité placentaire: recherche de biomarqueurs épigénomiques d'exposition au Benzo[a]pyrène par spectrométrie de Masse

R. Bilgraer, S. Gillet, O. Laprévote

- 3e Journée Scientifique de l'IMTCE, Paris 22 Juin 2012 **(Poster)**

Toxicité placentaire: recherche de biomarqueurs épigénomiques d'exposition au Benzo[a]pyrène par spectrométrie de Masse

R. Bilgraer, S. Gillet, L. Fernandes, J. Badet, S. Gil, D. Evain-Brion, O. Laprévote

Publication :

Raphaël Bilgraer, Sylvie Gillet, Sophie Gil, Danièle Evain-Brion, Olivier Laprévote (2014)

A new approach combining LC-MS and multivariate statistical analysis for revealing changes in histone modification levels

Mol. BioSyst. 10, 2974-2983

Résumé :

En influençant le degré de compaction de la chromatine ainsi que ses interactions avec différents partenaires protéiques, les modifications post-traductionnelles des histones sont impliquées dans la régulation de l'expression des gènes. Avec les différents variants d'histones incorporés dans la chromatine, ces modifications dynamiques et sensibles à l'environnement sont constitutives du code histone. Ce travail présente une approche globale de criblage baptisée approche histonomique, visant à révéler une perturbation épigénétique à l'échelle des histones. Cette approche originale offre une comparaison rapide et fiable des abondances relatives des variants d'histones et de leurs modifications post-traductionnelles dans des cellules humaines en une seule analyse LC-MS. Comme preuve de concept, des cellules BeWo issues de choriocarcinome humain ont été exposées au butyrate de sodium, un inhibiteur non spécifique d'histones désacétylases. Les histones extraites des échantillons témoins ou traités au butyrate de sodium à 1 ou 2,5 mM ont été analysées par chromatographie liquide ultra performante couplée à un spectromètre de masse de type Q-TOF. Les analyses statistiques multivariées ont permis de discriminer les échantillons témoins des échantillons traités sur la base des différences de degrés d'acétylation observés sur plusieurs formes d'histones. La même approche a ensuite été appliquée à des cellules exposées au B[*a*]P à 1 µM et a révélé deux principaux marqueurs caractéristiques d'un remodelage de la chromatine induit par les effets génotoxiques du B[*a*]P. En résumé, cette approche histonomique globale pourrait se révéler être un outil complémentaire très utile pour explorer une potentielle perturbation du code histone lors d'exposition à des xénobiotiques environnementaux.

Mots-clés : placenta, toxicologie, épigénétique, histones, modifications post-traductionnelles, chromatographique liquide, spectrométrie de masse, polluants environnementaux.

Abstract :

While acting upon chromatin compaction, histone post-translational modifications (PTMs) are involved in modulating gene expression through histone–DNA affinity and protein–protein interactions. These dynamic and environment-sensitive modifications are constitutive of the histone code that reflects the transient transcriptional state of the chromatin. Here we describe a global screening approach for revealing epigenetic disruption at the histone level. This original approach enables fast and reliable relative abundance comparison of histone PTMs and variants in human cells within a single LC-MS experiment. As a proof of concept, we exposed BeWo human choriocarcinoma cells to sodium butyrate (SB), a universal histone deacetylase (HDAC) inhibitor. Histone acid-extracts equally representing 3 distinct classes, Control, 1 mM and 2.5 mM SB, were analyzed using ultra-performance liquid chromatography coupled with a hybrid quadrupole time-of-flight mass spectrometer (UPLC-QTOF-MS). Multivariate statistics allowed us to discriminate control from treated samples based on differences in the acetylation level of several histone forms. We then applied the same procedure to cells treated with 1 µM B[*a*]P and suceeded in revealing two markers of chromatin remodeling in relation with genotoxic properties of B[*a*]P. Indeed, this untargeted histonomic approach could be a useful exploratory tool in many cases of environmental xenobiotic exposure when histone code disruption is suspected.

Key words : placenta, toxicology, epigenetics, histones, post-translational modifications, liquid chromatography, mass spectrometry, environmental pollutants.